SHANLAX

INTERNATIONAL JOURNAL OF ARTS, SCIENCE AND HUMANITIES

(A Peer-Reviewed-Refereed/Scholarly Quarterly Journal Globally Indexed with Impact Factor)

Vol. 5 Special Issue 2 February, 2018 Impact Factor: 2.114

ISSN: 2321-788X UGC Approval No: 43960

National Seminar on

EMERGING TRENDS IN COMPUTING TECHNOLOGIES

DEPARTMENT OF COMPUTER SCIENCE & INFORMATION TECHNOLOGY

Friday, 16th February 2018

MADURAI SIVAKASI NADARS PIONEER MEENAKSHI WOMEN'S COLLEGE

(Accredited with 'B' Grade by NAAC)

(Affiliated to Alagappa University, Karaikudi)

Poovanthi, Sivagangai District

EDITORS OF SPECIAL ISSUE JOURNAL

Dr. S. NagaNandhini Sujatha
Assistant Professor
Computer Science, Government Arts College for Women, Nilakottai

Ms. K. Sudharani
Head of the Department
Computer Science, Madurai Sivakasi Nadars Pioneer Meenakshi Women's College Poovanthi

Ms. K. Mahalakshmi
Head of the Department
Information Technology, Madurai Sivakasi Nadars Pioneer Meenakshi Women's College, Poovanthi

Ms. T. Ramaporkalai
Assistant Professor
Computer Science, Madurai Sivakasi Nadars Pioneer Meenakshi Women's College Poovanthi

Ms. P. Priya
Assistant Professor
Computer Science, Madurai Sivakasi Nadars Pioneer Meenakshi Women's College Poovanthi

Ms. N. Vinothini
Assistant Professor
Information Technology, Madurai Sivakasi Nadars Pioneer Meenakshi Women's College, Poovanthi

Ms. K. Sankareswari
Assistant Professor
Information Technology, Madurai Sivakasi Nadars Pioneer Meenakshi Women's College, Poovanthi

CONTENTS

GUIDELINE VALUE REGISTRATION

A.Dharmarajan
Assistant Professor, Ayya Nadar Janaki Ammal College, Sivakasi, India

P.Kavitha
Student of PG CS, Ayya Nadar Janaki Ammal College, Sivakasi, India

Abstract
The core objective of this mobile application is to create a convenient way to know about the cost of land or building at any time. The Guideline Values are already available in the website (www.tnreginet.net). Any user can access the details of site and assert values as per government norms. The app has multiple options and registration rules as per government regulation. This application acts like an interface between the people and government. Intermediate agents can't cheat buyer, seller and government anymore because this application provides the updated guideline Value.
Keywords: *Android, Eclipse IDE, MySql Database, XAMPP Server, PHP, Java*

Introduction

Android is an Operating system for smart phones. The Android OS is an open source operating system primarily used in java and based on the Linux Operating system. The Android OS consists of numerous Java applications and Java core libraries running under the Java – based object oriented applications framework and the Dalvik Virtual Machine. Dalvik is integral for the android to run in mobile devices as these systems are constrained in terms of processor speed and memory.

Eclipse SDK 4.2 is upwards workspace-compatible with earlier 3.x and 4.x versions of the Eclipse SDK unless noted. This means that workspaces and projects created with Eclipse SDK 4.1, 4.0, 3.8. 3.0 can be successfully opened by Eclipse SDK 4.2 and upgraded to a 4.2 workspace. This includes both hidden metadata, which is localized to a particular workspace, as well as metadata files found within a workspace project, which may propagate between workspaces via file copying or team repositories. Individual plug-ins developed for Eclipse SDK 4.2 should provide similar upwards compatibility for their hidden and visible workspace metadata created by earlier versions; 4.2 plug-in developers are responsible for ensuring that their plug-ins recognize metadata from earlier versions and process it appropriately. User interface session state may be discarded when a workspace is upgraded. Downward workspace compatibility is not supported. A workspace created (or opened) by a product based on Eclipse 4.2 will be unusable with a product based on an earlier version of Eclipse. Visible metadata files created (or overwritten) by Eclipse 4.2 will generally be unusable with earlier versions of Eclipse

Materials and Methods

PHP stands for Hypertext Preprocessor (it's a recursive acronym). PHP is a language used to develop interactive and dynamic content on the web and it is often used together with the Apache web server. It can also be used with Microsoft's IIS web server. PHP is a scripting language - Code written in PHP is not executed but interpreted by another program at runtime instead of being compiled by a computers processor. Other scripting languages include JavaScript and VBScript. PHP is a web language - It is used to create content on webpages. PHP is a server-side language - Unlike some other web languages, PHP is server-side as opposed to client-side. If you try to view the source code of a web page,

you will see that it may have been written in various language including HTML and Javascript. The source code you will be seeing however is client-side code - the code of those languages that execute on the browser. If a webpage is written using PHP, you will not see the PHP source code used to create the page if you try to view the source code of the page. While PHP executes on the server, and not the browser, PHP files will be returned to the browser as plain HTML. PHP is an object-oriented language - In PHP you can define your own reusable data structures called objects as well as define their attributes (properties) and things they can do (methods). You can also create relationships between various objects and data structures. PHP is an open source language - PHP's source code (the source code used to create PHP) is freely available to the general public.

JAVA

Java programming language was originally developed by Sun Microsystems, which was initiated by James Gosling and released in 1995 as core component of Sun Microsystems. Java platform (Java 1.0 [J2SE]). As of December 08 the latest release of the Java Standard Edition is 6 (J2SE). With the advancement of Java and its wide spread popularity, multiple configurations were built to suite various types of platforms. Ex: J2EE for Enterprise Applications, J2ME for Mobile Applications. Sun Microsystems has renamed the new J2 versions as Java SE, Java EE and Java ME respectively. Java is guaranteed to be **Write Once, Run Anywhere** Java is: In java everything is an Object. Java can be easily extended since it is based on the Object model. Unlike many other programming languages including C and C++ when Java is compiled, it is not compiled into platform specific machine, rather into platform independent byte code. This byte code is distributed over the web and interpreted by virtual Machine (JVM) on whichever platform it is being run. Java is designed to be easy to learn. If you understand the basic concept of OOP java would be easy to master. With Java.s secure feature it enables to develop virus-free, tamper-free systems. Authentication techniques are based on public-key encryption. Java compiler generates an architecture-neutral object file format which makes the compiled code to be executable on many processors, with the presence Java runtime system. Being architectural neutral and having no implementation dependent aspects of the specification makes Java portable. Compiler and Java is written in ANSI C with a clean portability boundary which is a POSIX subset. Java makes an effort to eliminate error prone situations by emphasizing mainly on compile time error checking and runtime checking. With Java.s multi-threaded feature it is possible to write programs that can do many tasks simultaneously. This design feature allows developers to construct smoothly running interactive applications. Java byte code is translated on the fly to native machine instructions and is not stored anywhere. The development process is more rapid and analytical since the linking is an incremental and light weight process. With the use of Just-In-Time compilers Java enables high performance. Java is designed for the distributed environment of the internet. Java is considered to be more dynamic than C or C++ since it is designed to adapt to an evolving environment. Java programs can carry extensive amount of run-time information can be used to verify and resolve accesses to objects on run-time.

Database Description

Database is the heart of any information system. It is a collection of information related to a particular object or purpose. Tables are Database objects to store data. Collection of fields is called tables. Collection of records is called fields. The general objective is to make information access easy, quick inexpensive and flexible for the users. In a database environment, common data are available a used by several users. The Input data set attributes and description are shown in Tables.

Column Name	Data Type	Description
Id	Number	Plinth Area Id (Primary Key)
Plinth Area	Alphabets	Name of a Roofing

Table 1 Description of the Zone Details

Column Name	Data Type	Description
Id	Number	Survey Number Id (Primary Key)
Zone	Alphabets	Name of a Zone

Table 2 Description of the Sub Register Details

Column Name	Data Type	Description
Id	Number	Survey Number Id (Primary Key)
Sub Register	Alphabets	Name of a Sub Register
ZID	Number	Zone Id

Table 3 Description of the Village Details

Column Name	Data Type	Description
Id	Number	Survey Number Id(Primary Key)
Sub Register	Alphabets	Name of a Sub Register
SID	Number	Sub Register Id Id

Table 4 Description of the Plinth Area Details

Column Name	Data Type	Description
Id	Number	Survey Number Id (Primary Key)
Survey Number	Alphabets	No of Sub divided Survey Number
Guideline Value	Number	Value of a land
Guideline Value in Metric	Number	Value of a land in Metric

Table 5 Description of the Survey Number Details

Column Name	Data Type	Description
Id	Number	Roofing Id (Primary Key)
Roofing	Alphabets	Name of a Roofing
RA	Number	Roofing Amount
PID	Number	Plinth Area Id

Table 6 Description of the Roofing Details

Results

The Project has been developed by using Android as front end and MYSQL as back end. I already mentioned the details of app in objective section. Guideline values will be displayed after user choose the street name or survey number. Then users have to submit the appropriate details about property or assert values such as: the type of assert building or land, roofing, flooring, dimensions, year of construction. After that the submitted values are calculated and valuation report will be displayed. The Guideline value will be updated and maintained periodically by the admin. My app will run on both mobile and laptop so my project has responsive view.

Results of Zone

Zone details can be inserted or updated or deleted by admin.

Figure 1: Zone

Results of Sub Register

When user selects their Zone and Sub Register can be inserted or updated or deleted by admin

Figure 2: Sub Register

Results of Village

When user selects their Zone & Sub Register Villages can be inserted or updated or deleted by admin

Figure 3: Village

Results of Survey Number

Zone, sub register and village will be displayed. When the user selects their village survey numbers can be inserted or updated or deleted by admin.

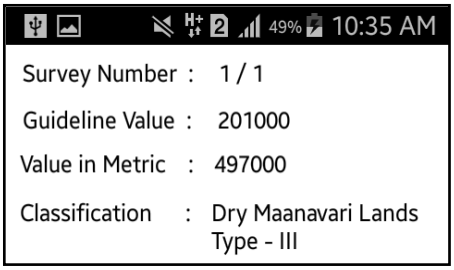


Figure 4: Survey Number

Results of Roofing

Type of assert building or land, roofing, flooring, dimensions, year of construction, EB Connection, Water connection amount will be modified by admin.

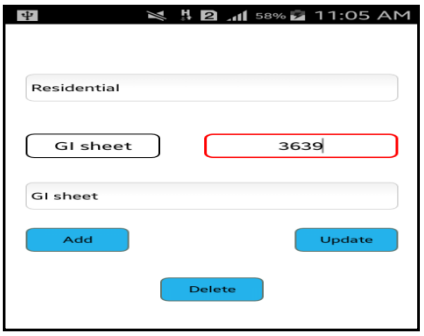


Figure 5: Roofing

Results of Report

After user choose the street name or survey number the guideline value will be displayed. In this module user have to input land extent, building extent, roofing, flooring, etc., to calculate their assert value. Building value will be generated and displayed.

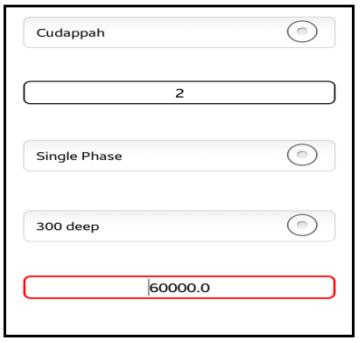


Figure 6: Report

Conclusion

Traditionally, people used to find their assert value by going to the register office or consult panel valuer. The implementation of Guideline value app digitalized this manual work. People can use this app any where any time to find their assert value. This system gives reliable secured guideline values. The values are periodically updated by the admin. This implementation done for a single zone (Madurai). In future, the project would be extended to whole all zones of Tamilnadu. Further, I am going to design this app not to find assert value only but to generate the Patta and Encumbrance certificate also. In Further, I will try to design this app is in Tamil language.

References

1. Richard Fairly, "Software engineering concepts", 1st Edition, Tata McGraw – Hill publishing company limited, New Delhi, 1997 [pp.123-150]
2. Marko G., "Learning android", 2nd Edition, O'Reilly Medi Inc., Sebastopol, March 2011 [pp.17-45, 78-92, 133-198, 204-241]
3. Kevin T., Peter MacIntyre, Rasmus Lerdorf, "Programming PHP", 3rd Edition, O' Reilly Medi, Inc., Sebastopol, February 2013 [pp.142-168, 246-285]
4. Michael Hernandez, "SQL Queries for Mere Mortals" 3rd Edition, A Hands-On Guide to Data Manipulation in SQL Publications, New Delhi, 1999 [pp.28-59, 67-82, 90-105]
5. http://www.tnreginet.net/GuidelineValue_2014/greportsro_2017.asp?zn=4.
6. https://www.javatpoint.com/android-tutorial.
7. https://www.youtube.com/watch?v=EknEIzswvC0&list=PLS1QulWo1RIbb1cYyzZpLFCKvdYV_yJ-E.
8. https://stackoverflow.com/questions/8078654/dynamic-multiple-spinners-state-city#
9. https://www.youtube.com/watch?v=EknEIzswvC0&list=PLS1QulWo1RIbb1cYyzZpLFCKvdYV_yJ-E.
10. http://www.geeks.gallery/category/android/ .
11. http://www.geeks.gallery/category/android/.
12. https://www.tutorialspoint.com/android/android_user_interface_controls.htm.
13. http://www.uml-diagrams.org/use-case-diagrams.html.

COLLEGE BUS PASS BOOKING AND TRACKING LOCATION APP

A.Dharmarajan
Assistant Professor, Ayya Nadar Janaki Ammal College Sivakasi, India

K.Selvakala
Student of PG CS, Ayya Nadar Janaki Ammal College Sivakasi, India

Abstract
This app helps both admin and students for bus pass automation. Students can apply through this app and also monitor the location of the college bus. The student can enter the details of the destination which will calculate the travel fare. The payment of bus fee can be processed by using paytm. The admin can monitor the location of the bus by capturing the longitude and latitude from the client to the database. The admin app will capture the location details and viewed through google map.
Keywords: *GPS, Google Maps*

Introduction

Android has become very popular in the world since it is an open source. In todays world, the time is more important for students. Being a product of high technology, mobile phones are more widely used and are becoming more and more popular. A vehicle tracking system is a commonly used application for tracking vehicles. Due to traffic congestion and road works, most of the buses are delayed. People have to wait for their bus at the bus stops for a long time without even knowing when the bus will arrive. Thus, the arrival time of the bus cannot be guaranteed. The main focus of the project is to save the waiting time of students and provide them the details of the bus.

Android Studio 2.0.0

The main advantage of adopting Android is that it offers a unified approach to application development. Developers need only develop for Android, and their applications should be able to run on numerous different devices, as long as the devices are powered using Android. In the world of smart phones, applications are the most important part of the success chain.

Android studio is the official integrated development environment (IDE) for Google's Android operating system, built on JetBrains' IntelliJ IDEA software and designed specifically for Android development. It is available for download on Windows, macOS and Linux based operating systems.It is a replacement for the Eclipse Android Development Tools (ADT) as primary IDE for native Android application development.

Features of Android Studio.

- **Connectivity**- Supports GSM/EDGE, IDEN, CDMA, UMTS, Bluetooth, Wi-Fi, LTE, and WiMAX
- **Messaging** - Supports both SMS and MMS.
- **Web browser -** Based on the open source WebKit, together with Chrome's V8 JavaScriptengine
- **Media support** - Includes support for the following media: H.263, H.264 (in 3GP or MP4container), MPEG-4 SP, AMR, AMR-WB (in 3GP container), AAC, HE-AAC (in MP4 or3GP container), MP3, MIDI, Ogg Vorbis, WAV, JPEG, PNG, GIF, BMP

- **Hardware support** - Accelerometer Sensor, Camera, Digital Compass, Proximity Sensorand GPS
- **Multi-touch -** Supports multi-touch screens
- **Multi-tasking -** Supports multi-tasking applications
- **Tethering** - Supports sharing of Internet connections as a wired/wireless hotspot

Components of Android Studio 2.0.0

Activity

- An activity represents the graphical user interface.
- It consists of buttons or other UI objects.
- The layout can be created in XML or dynamically in Java code

Services

- Services don't have a user interface
- They are used for working on tasks in the background.
- For instance, downloading some files or playing music

Content Provider

- This is used to exchange information between apps

Tracking GPS in Android

GPS stands for Global positioning system has wide number of application today popularly in the field of navigation, tracking etc. A GPS is a space-based navigation system that provides location and time information in all weather conditions, anywhere on or near the Earth where there is an unobstructed line of sight to four or more GPS satellites.

1. **Create a Location Listener** object, and implement the callback methods.
   - Location Listener loc Listener=new **Location Listener** () {
   - public void **on Location Changed**(Location location) {}
   - public void on Status Changed(String provider, int status, Bundle extras) {}
   - public void on Provider Enabled(String provider) {}
   - public void on Provider Disabled(String provider) {}

2. **Get a reference** to the Location Manager (system service)
   - Location Managerlm = (Location Manager) get System Service (Context. LOCATION_SERVICE)

3. **Register the LocationListener** in order to receive location updates from the Location Manager. lm. requestLocationUpdates(provider,minTime,minDistance,locListener)
4. **Add** user permissions in the **XML Manifest**

<manifest><uses-permissions android:name="android.permission.ACCESS_FINE_LOCATION" />
<uses-permissions android:name="android.permission.ACCESS_COARSE_LOCATION" />
<uses-permissions android:name="android.permission.INTERNET" />
</manifest>

System Analysis

Problem Definition

In my findings, the students need to apply bus pass in traditional and thus it will take lot of time and the process won't be user friendly. If the student missed the bus pass, it become a risky task to get a duplicate copy. Attaching GPS device to every college bus will be a risky task. The admin of the college bus can't view the current location and status of the bus, thus knowing the status of the college bus will become a risky factor.

Disadvantage of Existing System

- If the bus stops due to fault and accident, the admin can't get intimation in quick time.
- The amount spent for attaching GPS in bus will be more expensive.
- If any fault occurs in GPS device, it can't be overcome in quick time therefore capturing of location details may stop.
- The time consumption in maintaining bus pass stop details manually will take more time.
- All category of students can't make use of the bus pass and applying for duplication copy needs lot of procedure.
- Inaccurate location

Proposed System

To overcome the disadvantages in the existing method, it is necessary that we should use the college bus pass booking and location tracking application.

The user can apply their bus pass through app and as per the source and destination, the amount get predicted and the payment will be extracted from the wallet.

Merits of Proposed System

- The location information can significantly reduce the time it takes to access real-time arrival information for a nearby stop.
- The application can be used by the driver at any route which can be differentiate as per the login, therefore the unique id helps to know the status of the respective bus.
- The payment can be done online via paytm.

Module Description

- Admin Module
- Student Module
- Bus App Module

Administration

This module is for the bus administrator for updating the information that is there in the server when required. It includes authority to update Driver name, Driver Contact Number, Route, Stops, etc. The administrator needs to log in before editing or updating details. Only administrator is the authorized user of this module. The administrator provided that his internet connection is in enabled mode.

View Current Location

Student and staff must make sure that their location service is active. They can also get the estimated time of arrival of bus at their respective stops. It helps them to manage their time and arrive at their stop at the proper time, neither too early nor late.

View Student Details

After providing all the necessary details such as student proofs, college name, phone number etc ,the details will be submitted to the admin. The admin verifies and authenticates the detail. The admin can only able to view the details like Student's roll no, name, class and other details etc through the successful login.

Payment Using Paytm

The fare for the bus pass can be processed by using paytm. The wallet amount of paytm can be used to make the payment for bus pass.

Steps in Interacting with Paytm

1. Download PGSDK library from http://paywithpaytm.com/developer/#web_plugin by entering registered paytm mobile number.
2. Add the INTERNET and ACCESS_NETWORK_STATE permissions to AndroidManifest.xml. These two permissions are required for PGSDK Service to run in AndroidManifest.xml
3. <uses-permission android:name="android.permission.INTERNET"/>
4. <uses-permission android:name="android.permission.ACCESS_NETWORK_STATE"/
5. Add PaytmPGActivity in AndroidManifest.xml and this activity already present in PGSDK. This Activity perform all transaction related information.

UML Diagram

Activity Diagram

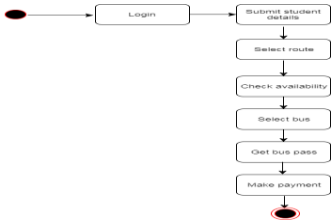


Figure 1 Activity diagram for Bus Pass

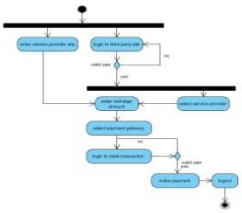


Figure 2 Activity Figure diagram for payment

Screenshots

Figure 3 Student Page

Figure 4 Student Page

Figure 5 Using paytm

Figure 6 Using Wallet

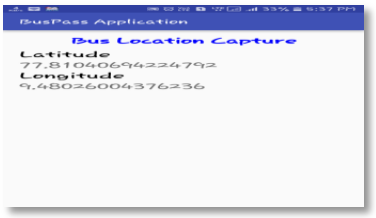


Figure 7 Location capture

Conclusion

This app is convenient for the students who are facing problems with the current manual work of college bus pass booking. It makes the students easy to travel with the bus pass with the mobile. So that even if the students loses the bus pass at the time of checking they can get the duplicate pass. This app is composed of smart phones and a server. The app is able to demonstrate its performance to track college bus from any area. It also estimates the time required to reach a particular stop on its route. This application uses client-server technology. Hence it is user-friendly and flexible to use as it is a time saving application to the user.

References

1. David Powers, "PHP Solutions Dynamic Web Design Made Easy", 1st Edition, APress, New York, 2010, [pp. 283 - 301]
2. W.Jason Gilmore, "Beginning PHP and MySQL", 4th Edition, Apress, New York, 2010, [pp 587 – 601]
3. Jeff Friesen, “Learn Java for Android Development” 3rd Edition, APress, New York, 2014, [pp 31 – 87]
4. Marko Gargenta, “Learning android”, 1st Edition, O’Reilly Media, Inc., Sebastopol, March 2011, [pp 47 – 82]
5. Adam Gerber, Clifton Craig, "Learn Android Studio" 1st Edition, Apress, New York, 2015, [pp 57 – 87]
6. www.java2s.com/OpenSource/Android_Free_Code/Map/Route/index.htm[July.29, 2017]
7. www.androidjson.com/android-php-insert-display-select-update-delete/
8. www.android-examples.com/get-selected-item-position-of-spinner-in-android / [Aug.29, 2017]
9. www.codeproject.com/Questions/1110034/Add-attachment-to-gmail-app-via-intent-android [Sep.3,2017]
10. www.androidhub4you.com/2013/09/send-email-with-attachment-in-android.html [Sep.8,2017]
11. www.itcuties.com/android/pick-image-from-gallery/ [Sep 12,2017]
12. www.developer.android.com/reference/android/content/Intent.html[Sep 15,2017]
13. www.w3schools.com/sql/sql_func_max.asp [Sep 20, 2017]
14. www.stackoverflow.com/questions/2789276/android-get-real-path-by-uri-getpath [Sep 25, 2017]
15. www.youtube.com/watch?v=LMN7iweea8c [Sep 27, 2017]
16. *www.apress.com/us/book/9781430231561 [Sep 30, 2017]*
17. www.youtube.com/watch?v=qS1E-Vrk60E [Oct 2, 2017]
18. www.youtube.com/watch?v=RUqXme0jx6w [Oct 5, 2017]

IMPROVED DATA SECURITY IN QR CODE

K.P.Maheswari & S.Nirmala Devi
Assistant Professors, Department of CS & IT
Subbalakshmi Lakshmipathy College of Science, Madurai

Abstract

The data security from unauthorized use is becoming a big challenge today. Even though, many cryptographic methods and algorithms evolved day by day complete security especially working online is hard to achieve [2]. One of the efficient way in protecting the data is through passwords, generating OTP, QR code, etc which may protect our identities on websites, email accounts and more in efficient manner. In this paper, we are going to propose a new improved way of data security by considering the password in QR code [11].

Data security is much more essential from its core that is from its storage place with the advancements in technology. QR (Quick Response) code is used to store data with higher / large storage capacity. The data embedded in QR code are publicly accessible. It needs more security since there is no access restriction, which in turn with the help of malwares leads to security attacks and also the data becomes insecure.

The possible security features for the data in a QR code are experimented in this paper. The data analysis is made to implement these concepts in real world scenario. Even unauthorized users(Hackers/Intrudes) hacks the QR code, the ways to block access of data embedded were also focused here. This may act as an efficient way of providing data security.

Keywords: *Data Security, Password, QR Code, File Protection.*

Introduction

Security plays a vital role in today's world. Many techniques such as Cryptography, Steganography and Watermarking can be used to obtain security, secrecy, privacy and authenticity of data. The process of data encryption and hiding[1][2][3] is achieved using QR code(Quick Response codes).QR code is a two dimensional bar code capable of encoding different types of data like binary, numeric, alphanumeric etc. Any information can be embedded into the QR code and it can be accessed in general public. It is also possible to use it for file protection.

Data Security

Data security refers to protective digital privacy measures that are applied to prevent unauthorized access to computers, databases and websites. Data security also protects data from corruption. Data security is an essential aspect of IT for organizations of every size and type.

Examples of data security technologies include backups, data masking and data erasure. A key data security technology measure is encryption, where digital data, software/hardware, and hard drives are encrypted and therefore rendered unreadable to unauthorized users and hackers.

One of the most commonly encountered methods of practicing data security is the use of authentication. With authentication, users must provide a password, code, biometric data, or some other form of data to verify identity before access to a system or data is granted.

QR CODE

QR Codes are split into various sections and it's these sections that the scanners use to decode the data. The image below is a structure of the QR code with various sections.

- ***Finder Pattern:*** The finder pattern consists of three identical structures that are located in all corners of the QR Code except the bottom right corner. The Finder Patterns enable the decoder software to recognize the QR Code and determine the correct orientation.
- ***Separator:*** The white separators have a width of one pixel and improve the recognizability of the Finder.
- ***Timing Pattern:*** Alternating black and white modules in the Timing Pattern enable the decoder software to determine the width of a single module.
- ***Format Information:*** The Formation Information section consists of 15 bits next to the separators and stores information about the error correction level of the QR Code and the chosen masking pattern.
- ***Encoding Region:*** Data is encoded into a bit stream and then stored in 8 bit parts (called code words) in the data section.

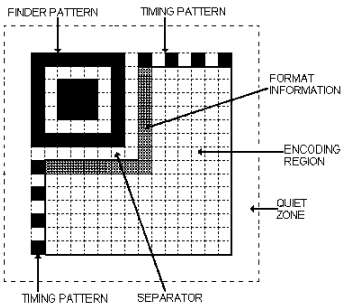


Figure 1: Structure of QR code

Proposed Work

In this approach, the security of the data is analyzed in terms of data loss by reducing the file size. By considering the two important factors such as file size and Security, the QR code is generated accordingly.

The security factor is implemented by taking two different files, such as one with password and another without password. Those files with different size are embedded in QR code and analyzed for data loss in the proposed algorithm.

Proposed Algorithm

Step 1: Create a PDF file with password.
Step 2: Generate QR code for the password protected PDF file.
Step 3: Store both QR code and PDF file in drive (online).
Step 4: Scan QR Code.
Step 5: Download the PDF by entering the password.

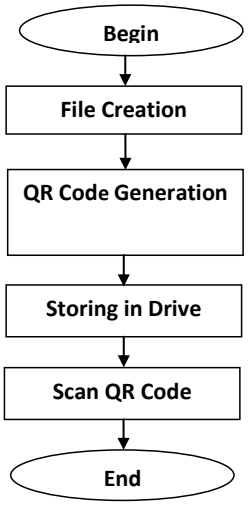


Figure 2: Flow Diagram of Algorithm

Work Analysis

In this proposed work, the study was made by analyzing the files stored in different forms such as PDF and QR code. It is observed that the size of the PDF is larger than the QR code which contains the same PDF file. The test was made with different file sizes including password.

By default, the data part in QR code patterns also differs for both PDF files with password and without password. The file size of QR code shows wide variation in size comparing to the original PDF.

QR Code – PDF 158 KB

QR Code – PDF 157 KB

File size		Data Protection
Pdf	Png – Qr code	
158 KB	95	With Password
157 KB	93	Without Password
3000 KB	95	With Password
2068 KB	94	Without Password

Table 1: Comparative Analysis

QR Code – PDF 3000 KB QR Code – PDF 2068 KB

Figure 3: Difference in data pattern of QR Code

When data in the form of PDF file is embedded in to the QR code, the variations in the size factors can be observed from the Table1

For file size less than 200 KB

- Difference in the size of the PDF files with and without password: 158 KB – 157 KB=1KB
- Difference in the size of the QR code for respective PDF: 95 KB–93 KB=2KB

For file size greater than 2000 KB

Difference in the size of the PDF files with and without password: 3000KB– 2068KB=932KB
Difference in the size of the QR code for respective PDF: 95 KB–94 KB=1 KB

Reduced Size

- 158 KB – 95 KB = 63 KB
- 3000KB – 95KB = 2905KB

The reduced size of files less than 200 KB is 63 KB and for files greater than 2000 KB is 2905 KB. The data of size more than 2000 KB is reduced to 1 KB, which shows the efficient way of data storage with security.

The reduction in the file size provides more spaces in drive and never the data loss occurs. In turn it reflects the improved way of data security in QR code.

Conclusion

This method of embedding the files in to QR code reduces the storage space and produced improved security. It could be used in large scope for file sharing, since the data access is possible without data loss. Also, it enables the user to store important data or information safely as QR code. The information could be retrieved easily from the QR code using QR reader.

It changes the convention of using QR code for public access in to improved and efficient storage of data with security.

References

1. P. Kieseberg, M. Leithner, M. Mulazzani, L. Munroe, S. Schrittwieser, M. Sinha, et al., "QR code security," in Proceedings of the 8th International Conference on Advances in Mobile Computing and Multimedia, 2010, pp. 430-435.
2. S. Dey, S. Agarwal, and A. Nath, "Confidential Encrypted Data Hiding and Retrieval Using QR Authentication System," in Communication Systems and Network Technologies (CSNT), 2013 International Conference on, 2013, pp. 512-517.
3. S. Dey, "SD-EQR: A New Technique To Use QR CodesTM in Cryptography," arXiv preprint arXiv:1205.4829, 2012.

4. D. Chatterjee, J. Nath, S. Dasgupta, and A. Nath, "A new Symmetric key Cryptography Algorithm using extended MSA method: DJSA symmetric key algorithm," in Communication Systems and Network Technologies (CSNT), 2011 International Conference on, 2011, pp. 89-94.
5. D. Sonawane, M. Upadhye, P. Bhogade, and S. Bajpai, "QR based Advanced authentication for all hardware platforms," International Journal of Scientific and Research Publications, vol. 4, pp. 1-4, 2014.
6. M. Bajpai and A. P. Agrawal, "INTEGRATION OF 2D SECURE BARCODE IN IDENTITY CARDS: WITH ADDITIONAL SECURITY FEATURES."
7. O. Sharaby, "Form Meets Function: Extreme Makeover QR Code Edition," ed: Archived from the original on, 2012.
8. H. Chan, "How to: Make your QR codes more beautiful," Maskable, April, vol. 18, 2011.
9. R.Cox.(2012).QArtCodes.Available:http://web.archive.org/web/20150321031237/http://researc.swtch.com/qart.
10. S. Hore, T. Bhattacharya, and S. B. Chaudhuri, "A Robust Medical Image Authentication Technique using QR Code and DWT," International Journal of Computer Applications, vol. 83, 2013.
11. Sawsan K. Thamer, Basheer N. Ameen, “A New Method for Ciphering a Message Using QR Code”. Published online at http://journal.sapub.org/computer Copyright © 2016 Scientific & Academic Publishing. All Rights Reserved
12. Ioannis Kapsalis., 2013. “Security of QR Codes“, Master thesis submitted in June 2013, Norwegian University of Science and Technology
13. Kinjal H. Pandyaand Hiren J. Galiyawala, 2014.”A Survey on QR Codes: in context of Research and Application”, International Journal of Emerging Technology and Advanced Engineering, Vol. 4 Issue 3.,pp. 258-262Ioannis Kapsalis., 2013. “Security of QR Codes“, Master thesis submitted in June 2013, Norwegian University of Science and Technology
14. QR Stuff QR Code Error Correction, 2011.QRStuff blog:http://www.qrstuff.Com
15. /blog/2011/12/14/QR-code-error-correction.

COMPARATIVE STUDY OF RANDOM FOREST AND LINK GUARD ALGORITHMS TO ENHANCE THE DETECTION OF PHISHING WEBSITES USING R

M.S.Najima Begum
III-B.Sc (Computer Science), Ayya Nadar Janaki Ammal College, Sivakasi

P.Tamizh Chelvi
Head of the Department of Computer Science,
Ayya Nadar Janaki Ammal College, Sivakasi

Abstract
Machine learning is a field within computer science, it differs from traditional computational approaches. Machine learning algorithms allow computers to train on data inputs and use statistical analysis in order to output values that fall within a specific range. Because of this, machine learning facilitates computers in building models from sample data in order to automate decision-making processes based on data inputs. Machine Learning has many numbers of algorithms. Here I am focusing on the comparative study of Random Forest and Link Guard Algorithms to enhance the Detection of Phishing websites using R. R is a Programming language and Software platform for Statistical analysis, graphics, representation and reporting Both algorithms have followed different ways and parameters to detect the phishing websites. They used many number of parameters to detect the websites exactly. Here, I'm trying to prove that the minimal parameters are sufficient for detection of phishing websites. By using those minimal parameters, we would be able to identify the phishing websites.
Keywords: *Link guard algorithm, Random access algorithm, Phishing.*

Introduction

Phishing is the third cyber-security threat globally and the first cyber-security threat in China. The word "Phishing" initially emerged in 1990s. The early hackers often use "ph" to replace "f" to produce new words in the hacker's community, since they usually hack by phones. Phishing is a new word produced from "fishing", it refers to the act that the attacker allure users to visit a faked Website by sending them faked e-mails (or instant messages), and stealthily get victim's personal information such as user name, password, and national security ID, etc. These informations then can be used for future target advertisements or even identity theft attacks. Machine Learning is one of the platforms that provides more algorithms and techniques to detect that type of phishing websites and e-mails. Machine Learning has many numbers of algorithms. Here, we are focusing on the comparative study of Random Forest and Link Guard Algorithms to enhance the Detection of Phishing websites using R. R is a Programming language and Software platform for Statistical analysis, graphics, representation and reporting. Both algorithms have followed different ways and parameters to detect the phishing websites. They used many number of parameters to detect the websites exactly. Here, we try to prove that the minimal parameters are sufficient for the detection of phishing websites. By using those minimal parameters, we would be able to identify the phishing websites.

Types of Methods Used with Machine Learning

- Decision Tree based Methods
- Linear Regression based Methods
- Neural Network
- Bayesian Network
- Support Vector Machine
- Nearest Neighbour

Here, we have selected **Decision Tree** based classification to classify the most Nearest attribute used to find out the Phishing website.

Random Forest	Link Guard
It is one of the classification methods	It is also one of the classification methods
The result accuracy of this algorithm is 99.7%	The result accuracy of this algorithm is 99.1%
It uses both low false negative (FN) and low false positive(FP) rates	It uses low false negative (FN) only.
To train the dataset, it uses Vector representation.	To train the dataset, it uses Pattern matching.
It uses regression	It uses end-host based approach

Figure 1 Comparison of Random forest and Link guard

Attributes used by RANDOM FOREST ALGORITHM to predict Phishing Websites:

- URL containing IP Addresses
- Presence of "Link," "Click," and "Here" in Link Text of a Link
- Number of Dots in Domain Name
- HTML Email
- Presence of Javascript
- Number of Links
- Number of Linked To Domain
- From_Body_MatchDomain Check
- Word List Features

User; Customer; Client; Suspend; Restrict; Hold; Verify; Account; Notify; Login; Username; Password; Click; Log; SSN; Social Security; Secur; Inconvinien

Attributes used by Link Guard Algorithm to predict Phishing Websites:

- Checking of Spoofed E-mails
- Actual and Physical Link
- DNS Address of Link
- Presence of Dotted domain name in the Link
- IP Address of the DNS Name
- Checking of Hyperlink Encoding

Here, the common attributes used in both algorithms are, DNS address, E-mail Verification, IP Address of the DNS Name, and Hyperlink encoding. Among these, we have to find the most important and the minimal parameters to classify the phishing websites.

Decision Tree

Decision tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label. Applying this algorithm, the most effective attribute can be found out to detect the phishing website. The leaf nodes are ignored because they are not needed for that.

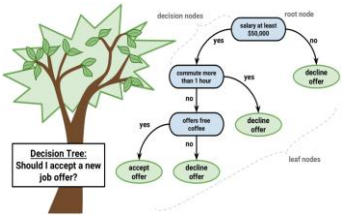


Figure 2: Decision Tree

Working of Decision tree Algorithm

- Place the best attribute of the dataset at the **root** of the tree.
- Split the training set into **subsets**. Subsets should be made in such a way that each subset contains data with the same value for an attribute.
- Repeat step 1 and step 2 on each subset until you find **leaf nodes** in all the branches of the tree.

Attribute Selection Measures in Decision Tree

The below are the some of the assumptions we make while using Decision tree:

- At the beginning, the whole training set is considered as the **root.**
- Feature values are preferred to be categorical. If the values are continuous then they are discretized prior to building the model.
- Records are **distributed recursively** on the basis of attribute values.
- Order to placing attributes as root or internal node of the tree is done by using some statistical approach.

The popular attribute selection measures

If dataset consists of **"n"** attributes then deciding which attribute to place at the root or at different levels of the tree as internal nodes is a complicated step. By just randomly selecting any node to be the root can't solve the issue. If we follow a random approach, it may give us bad results with low accuracy. For solving this attribute selection problem, researchers worked and devised some solutions. They suggested using some **criterion** like,

- Information gain
- Gini index

Here, we use Gini index for attribute selection. Using this, we are able to predict the minimal attributes needed to find out the phishing websites.

Conclusion

Phishing has becoming a serious network security problem, causing financial lose of billions of dollars to both consumers and e-commerce companies. And perhaps more fundamentally, phishing has made e-commerce distrusted and less attractive to normal consumers. In this paper, we have tried to find out the minimal parameter(s) needed to detect the phishing websites.

References

1. P. Prakash, M. Kumar, R. R. Kompella, and M. Gupta, "PhishNet: predictive blacklisting to detect phishing attacks," in Proceedings of the IEEE Conference on Computer Communications (IEEE INFOCOM '10), pp. 1–5, IEEE, San Diego, USA, March 2010..
2. UCI Machine Learning Repository." http://archive.ics.uci.edu/ml/, 2012.
3. Phishing Websites Creators." https://phish5.com/"
4. A. Bergholz, J. de Beer, S. Glahn, M. F. Moens, G. Paaß, and S. Strobel, "New filtering approaches for phishing email," Journal of Computer Security, vol. 18, no. 1, pp. 7–35, 2010.
5. R. Basnet, S. Mukkamala, and A. H. Sung, "Detection of phishing attacks: a machine learning approach," in Soft Computing Applications in Industry, pp.
6. StatisticalR. https://rdatamining.com"
7. Packages for Decision tree in RStudio. "https://cran.r-project.org/"

ARMY BORDER SECURITY SYSTEM

M.Sathya

III B.Sc(Computer Science), Ayya Nadar Janaki Ammal College, Sivakasi

R.Ananthavalli

Assistant Professor (Computer Science), Ayya Nadar Janaki Ammal College, Sivakasi

Abstract

For every country security and surveillance has become a key to survival. With rise in terrorism coupled with challenging security conditions, investment in security and surveillance by India is a necessity for survival. In the field of security and surveillance IoT based application can be utilized remotely to see the movement and get warnings when movement is identified. The photographs and recordings are sent straight forward to a cloud server, sent as Gmail Notifications with snapshots and SMS alerts for further action. Accordingly, points of interest such as these make IoT applications perfect for smart security surveillance monitoring wherever security is a big concern. Android Studio is the Integrated Development Environment to develop android applications. That applications can use with the embedded systems to establish the communication facilities. Here I would be creating an app to keep watch on the surveillance.

Keywords: *IoT, Android, RaspberryPi*

Introduction

In present days, all are very much aware about safety concerns. By knowing about the attacks done nearby our country border, we want to reduces the attacks by detecting the enemy using our system. This paper implements a prototype satellite based system to detect suspected people using Passive Infrared Sensor . This system also issues alert messages when any suspected human or non-human gets detected around border. The system consists of Raspberry pi acting as processing and controlling engine interfaced to GSM module for SMS alert, SD card where the input images are stored. We implement human detection algorithm using OpenCV. The application reads the input images, divide the image into several blocks for block based processing and search as for human in each block. The process is repeated at multiple scales so that we can detect humans in the input image at different sizes. Each block undergoes a sequence of machine learning operations like histogram computation, gradients calculation, binning process and finally classification using SVM. Finally, if any human or non-human is found in the input images, an alert message is send to army officials for further investigation.

Overview of the Proposed System

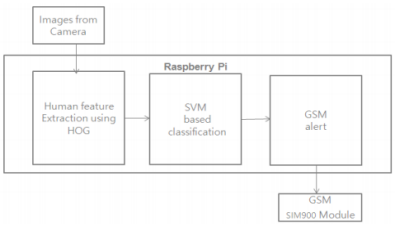


Figure 1: Block diagram

It clearly explains the whole proposed system. The main components are the Raspberry Pi processor, picamera and GSM module, Raspberry Pi consists of three sections viz., as Histogram Oriented Gradients(HOG), Support Vector Machine(SVM) and GSM SIM 900 Module. We installed Opencv library for the purpose of SVM and Wiring Pi for the purpose of GSM. The images taken from the camera are given as the inputs to the processor. The processed data will be send as SMS alert, if any suspected person detected at the border to convey to the army official about the situation for further investigation, gets clear description of the components in the system are given below:

The Raspberry Pi is a series of small single-board computers developed in the United Kingdom by the Raspberry Pi Foundation to promote the teaching of basic computer science in schools and in developing countries. The original models become far more popular than anticipated, selling outside of its target market for uses such as robotics. Peripherals (including keyboards) are not included with the Raspberry Pi. Some accessories however have been included in several official and unofficial bundles. LINUX is the operating system installed in Raspberry pi. Linux is a Unix-like computer operating system assembled under the model of free and open source software development and distribution.

Specifications

- CPU: Quad-core 64-bit ARM Cortex A53 clocked 1.2 GHz.
- GPU: 400MHz Video Core IV multimedia
- Memory: 1GB LPDDR2-900 SDRAM (i.e. 900MHz)
- USB ports: 4
- Video outputs: HDMI, composite video (PAL and NTSC) via 3.5 mm jack
- Network: 10/100Mbps Ethernet and 802.11n Wireless LAN
- Peripherals: 17 GPIO plus specific functions, and HAT ID bus
- Bluetooth: 4.1
- Power source: 5 V via Micro USB or GPIO header
- Size: 85.60mm × 56.5mm
- Weight: 45g.

Pin Diagram for Raspberry

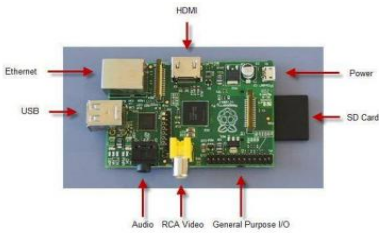


Figure 2: Pin diagram

PI Camera

A camera is an optical instrument for recording or capturing images, which may be stored locally and transmitted to another location, or both. The images may be individual still photographs or sequences of images constituting videos or movies. The camera consists of a small (25mm by 20mm by 9mm) circuit board, which connects to the Raspberry Pi Camera Serial Interface (CSI) bus connector via a flexible ribbon cable. The camera's image sensor has a clear resolution of five megapixels and has a fixed focus lens.

Hardware Implementation

It shows the hardware implementation of the system. Here the receiver pin and ground pin of GSM module is interfaced to the transmitter pin and ground pin of Raspberry Pi processor respectively .The Pi camera is connected to the Raspberry Pi .This Raspberry Pi process receives information from the inputs and gives necessary instructions to GSM Module.

Conclusion

By using this project we can save the lives of our army people. The Indian Army has forever been known for its courage and valor. They have protected us at times when our survival was doubtful, and have given their lives to save ours. Every Indian is proud of the Indian Army and we present this project as a salute to our Army men. Jai Hind!!

References

1. Ms. Sejal V. Gawande*, Dr. Prashant R. Deshmukh, "Raspberry Pi Technology", International Journal of Advanced Research in Computer Science and Software Engineering, Special Issue on Impact of Technology on Skill Development, Conference Held at IETE Amravati Center, Maharashtra, India, volume 5, issue 4,pp.37-40, April 2015.
2. [2] Rahul Antony, Reema Mathew A, Divya K, Teenu Jose, " Multi Security System Using GSM and PIC 16F877A", International Journal of Advanced Research in Computer Science and Software Engineering, volume 5, issue 3, pp.498-502, March 2015.
3. "Raspberry pi", available Https://www.raspberrypi.org/documentation/installation/noobs.md"
4. "Wiring pi library", available at :- "http://wiringpi.com/"
5. "installation of wiring pi" available at :- "To view the wiring Pi sources, then go to https://git.drogon.net/"

A STUDY ON BIG DATA IN CRISIS ADMINISTRATION

M.Saranya
M.Phil. Scholar, Department of Computer Science
Raja Doraisingam Government Arts College, Sivaganga, Tamil Nadu, India

A.Prema
Assistant Professor, Department of Computer Science
Raja Doraisingam Government Arts College, Sivaganga, TamilNadu, India

Abstract
Crisis Administration never run with earlier intimations and indications in the real world and practical life, detecting and perceiving some crisis. Large scale data with crisis administration helps to overcome the traditional problem having manual intercession and reporting emergency. Big data is the expert techniques and technologies to practice a very bulky set of data. These dataset are often so large and complex, it becomes difficult to process using on hand database management tools. Hadoop is an open source software framework and gratis java based programming structure that supports the processing of bulky dataset in a distributed computing environment. MapReduce is a programming model for computation massive amount of data and execution structure for large scale data processing. This paper presents the various methodologies used in crisis administration.
Keywords: *Crisis Administration, BigData, Hadoop, MapReduce.*

Introduction

Big Data is a group of data sets so huge and compound that it becomes difficult to route using on-hand database management tools . In Bigdata the data is generated from various different sources and can arrive in the system at various rates. Big Data into many dimensions: Volume, Velocity,Variety, Veracity and Value [14].

- **Volume**: The amount of data is at very large scale. The amount of information being collected is so huge that modern database management tools are unable to handle it and therefore become obsolete.
- **Velocity**: We are producing data at an exponential rate .It is growing continuously in terabytes and petabytes.
- **Variety:** We are creating data in all forms -unstructured, semi structured and structured data. This data is heterogeneous is nature. Most of our existing tools work over homogenous data, now we require new tools and techniques which can handle such a large scale heterogeneous data.
- **Veracity:** The data we are generating is uncertain in nature. It is hard to know which information is accurate and which is out of date.
- **Value**-The data we are working with is valuable for society or not.
- **Variability:** Variability in big data's context refers to a few different things. BigData is also variable because of the host of data dimensions resulting from multiple different data types and sources. Variability can refer to the erratic speed at which big data is loaded into your database.

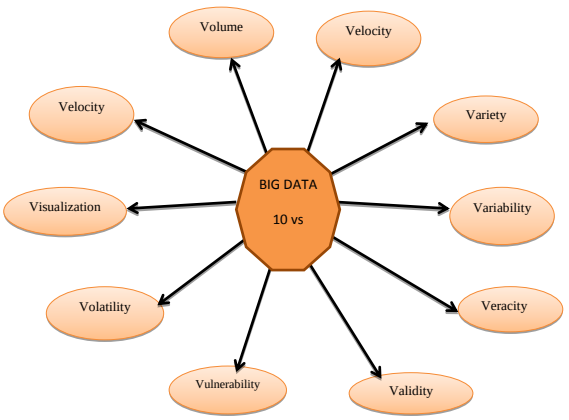


Figure 1: Big Data

- **Validity:** validity refers to how the data accurate and correct the data is for its intended use.

  Vulnerability: Bigdata brings new security concerns. The increasing size of people personal data, they have started feeling that it is being used to interfere into their behavior to sell them things by different commercial websites.
- **Volatility:** Especially in production systems, one has to prepare for data volatility. Data that should "never" be missing suddenly disappears, numbers suddenly contain characters.
- **Visualization:** Bigdata visualization tools face technical challenges due to limitations of in memory technology and poor scalability, functionality, and response time. We can't rely on traditional graphs when trying to plot a billion data points. So we need different ways of representing data such as data clustering or using tree maps, sunbursts, parallel coordinates, circular network diagrams, or cone tree.

Related Works

Manish Kumar Kakhani, SweetiKakhani, et al. describe the Big Data has attracted a lot of attention from academia, industry as well as government. Big Data is term defining collection of large and complex data sets that are difficult to process using conventional data processing tools. Big data is an emerging trend and there is immediate need of new machine learning and data techniques to analyze massive amount of data in near future [18].

Priyanka Rana says working Big databig data is not the amount of data the matters but the quality of information that can be extracted from database .In an organization Big data is evaluated for insights that direct to better strategic decisions.Advanced data analytics techniques like predictive analytics, location intelligence, and data mining are used to process hundreds of terabytes of data for financial decision making or business informatics [21].

Jyotikumari and Mr.Surender says the Bigdata analysis has seen an exponential growth and highly integrated system driven by rapid growth in information services and microelectronic devices. The current mobile systems are mainly targeted to voice communications with low transmission rates. Big data access at high transmission rate will be. Big data system that include a set of tools and technique to load , extract and improve dissimilar data while leveraging the immensely parallel processing power to perform complex transformations and analysis.[16]

Amir Gondomi and MurtazaHaider define analytics methods used for big data. They focus on analytics related to unstructured data, and highlight the need to develop appropriate and efficient analytical methods to leverage massive volumes of heterogeneous data in unstructured text, audio, and video formats. And also predictive analytics for structure big data.Bigdata analytical techniques for structure and unstructured data. Bigdata is namely text, audio, video and social media type. Big data potential value is unlocked only when leveraged to drive decision making. To enable such evidence-based decision making, organizations need efficient processes to turn high volumes of fast-moving and diverse data into meaningful insights..[3]

VarshaB.Bobaddescribes "Big Data" is a collected works of large data sets that cannot be processed using computer techniques. Big data is not just data somewhat it has become a whole subject which involves an assortment of the tools techniques , and frame work .the needed of big data generated by the large corporation like face book, Yahoo, Google.You tubeetc. for the purpose of analysis of huge amount of data also Google contains a large amount of information.[25]

Nawsherr Khan et al sayBig data is still in its infancy stage and the domain has not been reviewed in general. This reviewcatalogs the several elements of big data, including, nature, definitions rapid growth

rate, volume, management, analysis, and security Also proposed a data life cycle that uses the technologies and terminologies of big data. They facilitate the exploration of the domain and the development of optimal techniques to address big data.[20]

Kuchipudi Sravanthi and Tatireddysubba Reddy says the big data applications used in various fields .such as banking ,agriculture, chemistry , data mining, cloud computing, finance, marketing , stock, healthcare, etc. In this fields every fields has their concept and gave their usage related to big data.[17]

Anwaar Ali discuss the emerging ability to use big data techniques for development promises to revolutionize healthcare, education, and agriculture facilitate the alleviation of poverty and help to deal with humanitarian crises and violent conflicts. Beside all the benefits the large scale deployment of BD4D is beset with several challenges due to the massive size, fast changing and diverse nature of big data. Most of the BD4D project dealing with humanitarian emergencies is that they essential spring to action after the crisis has taken a huge toll. The real promise of BD4D is to use predictive analytics to avoid or mitigate such humanitarian emergencies before they can strike their toll. They used machine learning (ML) techniques is a sub field of artificial intelligence(AL) focuses on the enabling computational systems to learn from data how to perform a desired task automatically. Association Rule learning methods for discovering interesting relations between variables in large databases.[5]

G.suganya MCA designate nowadays the vast development of networking,data collection and data storage may lead to increase in size of data called as big data.Bigdata is the large volume of data that comes from various sources. it is a collection of larger data that cannot be processed using traditional database management system computing techniques. Bigdatavolume measured in terms of terabytes or petabytes. They proposed some techniques that may increase the performance in any time and in any situation. This may process not only text but also the other types of files format in an advanced manner.[24]

Chaitali Mohite ,Nisha Ambrate ,et al says the Hadoop has its own file system. Hadoop applications are most widely accessible .they discuss application of Hadoop ,the challenges of big data ,the technology of Hadoop, comparison of Hadoop architecture and its advantages and disadvantages.[6]

Ms.Rucha S Shelodkar, and Prof. Himanshu, proposed a modified MapReduce architecture that allows data to be pipelined between operators. The modified version of the Hadoop Map Reduce Framework that supports on the fly aggregation They proposed technique is to significantly improve the performance of HadoopMapreduce for efficient Big Data processing. Mapreduce supports on the fly aggregation and stream processing also improving utilization and reducing response time. Mapreducde on the fly technique can be used by allowing pipeline between operators with preserving fault – tolerance guarantees. This technique explains partial query processing and also user can observe the progress of running queries and control their execution. On the fly aggregation techniques to build tools for easy and efficient processing of very large data sets.[22]

Mr.A.AntonyPrakash and Dr.A.Aloysius discuss the Hadoop and MapReduce applications. Hadoop and MapReduce can be used for analyzing enormous amount of data. MapReduce is a programming model that associated implementation for parallel processing of large dataset .Bigdata environment is used to acquire,organize and analyze the various types of data. This framework generates large amount of intermediate data.[4]

Jeffery Dean and Sanjay Ghemawat describe the MapReduce runs on a large cluster of commodity machines and are highly scalable. Map Reduce programs all run in more number Mapreduce jobs.it executed on Google's clusters every day [15]

Chen xin, Cui Bin-ge et al., describea novel model for defining emergency plans, in which workflow segments appear as a constituent part. It contains four operations to define the workflow segments under the limit. The software system of the business process resource construction and composition is implemented and integrated into Emergency plan management application system. Emergency plans are a formal processing from documental emergency plan text to a structured data. In emergency information system, the emergency event is one of the key research objects. The information of event contains time, place, character, task, and consumption of resource. Which is 5Wprinciple: when, where, who, what, how. The pre-arranged emergency plans contain these 5 kinds of information, which is formulated by all kinds of emergency response units. The system mainly helps emergency decision-maker to schedule rescuers, their tasks, and emergency resources. Decision –maker needs a global view, which can get information simple and clear, and information contains responsible person rescue resource and task which are defined in plans. Emergency rescue departments maintain their plan data daily. Business process resource through a visual method to build, and decision makers can combine business process flexibly to provide decision support services.[7]

Christian Sell, and Iris Braun describe the model for workflow management system (WfMs) for supporting the modeling, execution and management of emergency plans before and during a disaster .It is based on the idea that emergency plans are similar to business processes and can therefore be modeled as workflows. In contrast to most traditional WFMS, the supports unstructured activities and their delegation as well as the management of resources The execution of emergency plans with and without the use of a WfMs with regard to time and effort reduction. [8]

A Mobaraki, A Mansourian et al discuss the immediate and time penetrating nature of trauma situations it is essential to collect and use spatial information of the present state of the emergency within the least waste of time. Spatial data infrastructure (SDI) is an appropriate framework to facilitate such collaboration in spatial data collection and in field decision making .Mobile GIS is a movable GIS that makes spatial data acquisition ,storage sharing analysis in every time and everywhere possible for users. Mobile GIS an efficient technology in managing spatial data, particularly in emergency management. Mobile GIS facilitates in field data collection and real time updating of EOC database collected routes, etc., Mobile GIS emergency workforces can access to EOC database which epitomizes present status of emergency situation and analyzing these data to make assessment for emergency operations. SDI conceptual model for emergency management to support mobile GIS was depicted.[19]

Aaron Malveaux and A Nicki Washington, describe the Emergency preparedness is a restraint that harness technology, citizens and government organizations to handle and potentially avoid natural disasters and emergencies. The survey of the use of information technology, including social media in emergency preparedness is presented. Social media can be leveraged to warn the public of emergencies and analyze how people respond to emergency situations. The web has served as a vessel filled with vital information , and therefore , has introduced a novel and unique aspect to emergency preparedness research information from social networks can be used to monitor events in real time, as well as gather information after emergencies. Mobile devices fortified with GPS hardware can collect this information. GPS receiver can find information that is appropriate to their location and surrounding .without the conveniences that these system offer, important information on events such as natural disasters or terrorist attacks would be lost to many people.[1]

DontasEmmanouil and DoukasNikolaos describe the Bigdata analytics techniques and tools that are useful in all phases of crisis management are presented .furthermore, a system engineering approach of big data management system will be analyzed that comprises of four phases, data generation ,data

acquisition ,data storage ,and data analytics. Data analysis, the goal of which is to extract useful values, suggest conclusions and /or support decision –making. Data analytics addresses information obtained through observation, measurement, or experiments about a phenomenon of interest. The aim of data analytics is to extract as much information as possible that is pertinent to the subject under consideration. In the big data a acquisition phase, typical data collection technologies were investigated during each stage of the data the life cycle management of big data is the most demanding issue.[9]

Adam Flizikowski et al describe the Social media communities has become a powerful and promising solution in crisis and emergency management. Crisis events proved that social media and mobile technologies used by citizens and public service have contributed to the post –crisis relief efforts. The crisis management especially in search and rescue operation. They focus on the end user's needs of access to data/information coming from social media channels, feature required platform and , main challenges for integrating social media into response efforts. They have showed that additional knowledge, training and guidelines on how to use social media for crisis management are expected by end –users.[2]

Bigdata and Hadoop

Big data value chain is data analysis, the goal of which is to get useful values, suggest best conclusion and support decision making system of an organization to stay in competition market.

- **Descriptive analytics:** Exploits historical data to describe what occurred in past. For instance, a regression technique may be used to find simple trends in the database sets, and data modeling is used to collect, store and cut the data in an efficient way. Describe analytics is typically associated with business intelligence or visibility system**.**
- **Predictive analytics**: Predictive modeling uses statistical techniques such as linear and logistics regression to understand trends and predict future outcomes, and data mining extracts patterns to provide insight and forecasts.
- **Prescriptive analytics**: Address decision making and efficiency. Analyze complex system to gain insight into system performance and identify issues and optimization techniques are used to find best solutions under given constrains.[16]
- **Technology for Bigdata:** BigData handle two classes of technology. Operational bigdata: Systems like Mongo DB that provide operational capabilities for real-time, interactive workloads, where data is primarily captured and stored .operational big data workloads much easier to manage, cheaper, and faster to implement. This based on minimal loading and without the need for data scientists and additional infrastructure.

Analytical bigdata: Systems like massively parallel processing database system and MapReduce . System based on Mapreduce that can be scaled up from single severs to thousands of high and low end machines [23].

Hadoop

Hadoop is the central platform for forming big data and solve the problem in useful method for analytics purpose. Hadoops an open source java framework technology is used to store, manage and distribute bigdata across several server nodes. It is archive distributed object oriented programming. Apache Hadoop comprise of a packing part and managing part (MapReduce). Hadoop splits files into large blocks and distributes them amongst the node in the cluster. To process the data, Hadoop

Mapreduce transfers packaged of data locality nodes to process in parallel, based on the data each nodes need to process.

The purpose of HDFS is to store large datasets reliably and to stream them at high bandwidth to user applications. HDFS has two types of nodes in the schema of master-worker pattern: a name node, the master and an arbitrary number of data nodes, the workers.

Name Node

The name node is the central location for information about the file system deployed in hadoop environment .An Environment can have one or two nodes, minimal redundancy between the name nodes. The name node is contacted by clients of the hadoop Dirstibuted file system (HDFS) to located information within the file system and provides updates for data they have added, moved, manipulated or deleted.

Data Node

Data node make up the majority of server contained in hadoop environment common Hadoop environment will have more than one data node .The data node serves two functions, it contains a segment of the data in the HDFS and it acts as a compute dais for running job, utilize narrow data within the HDFS.

Edge Node

The Edge node is the entrance point for the exterior applications tools, and users that need to utilize the Hadoop environment.[18]

Secondary Node

Secondary Name node acts as a backup for Name node. It stores the Meta data information from Name node at regular interval as per checkpoint mechanism.

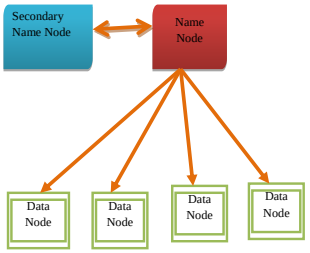


Figure 2 Hadoop Architecture [11]

Hadoop can route tremendously enormous volumes of data with varying structures. Hadoop serene of HBase, H catalog, pig, Hive, OOzie ,Zookeeper, and Kafka, however the most shared modules and well known paradigms are Hadoop Distributed File System(HDFS) and MapReduce for big data.

Hadoop Framework includes these Major Components

- **Hadoopcommon:**It is a collection of common utilities and libraries that support the other HadoopmoduelsHadoop common is also known as Hadoop core. It is an essential part of Hadoop frame work.
- **HDFS:**A distributed file system that provides high flow capacity access to application.HDFS used for storing all the structured data and unstructured data hence, it permits processing of the stored data.[6]
- **Hadoop Map Reduce:** a programming model for large scale data dealing out.
- **Cost effective:** Hadoop saves cost as it employs a cheaper low end cluster of commodity of machines instead of the costlier high end.

Map Reduce

MapReduce is a programming model for calculations on massive amounts of data and acompletingstructure for vast scale data processing on clusters of commodity servers. It was originally developed by Google and built on well-known principles in parallel and distributed processing [18]. MapReduce program consists of two functions –Map function and Reduce function.

1. Each Map function is converted to key-value pairs based on input data. The input to map function is tuple or document. The input data codes are written by the user for the Map function
2. The key-value pairs from each Map task are composed by a master controller and sorted by key. The keys are separated among all the Reduce tasks, so all key-value pairs with the same key wind up at the same Reduce task.
3. The Reduce tasks work on one key at a time, and syndicate all the values associated with that key in some way. the codes marked by the user for the Reduce function.

Mapreduce model hide details related to the data storage distribution replication load balancing and so no. Map function and reduce information for performing the processing of the bigdata.

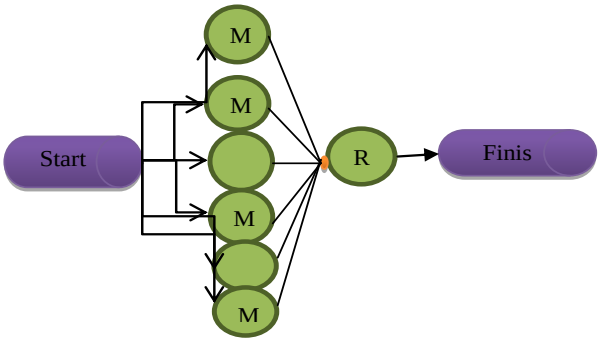


Figure 3 : MapReduce [12]

Crisis Administration

Crisis administration is the practice by which an organization deals with a trouble and astonishing event that threatens to harm the organization. Its stakeholders or the general public .crisis management is a situation based management system that includes clear role and responsible and process related organization Requirements Company wide. The Crisis is a condition characterized by surprise a high risk of serious values and short reaction time ".The four phases of crisis Preparedness, Response and Recovery.

Big Data and Crisis Prevention

Information derived from the analysis of Big Data can help to anticipate crises or a least reduce the risk that would arise from a disaster the major crisis effect .One example is in a big earthquake harm arises in telecommunication networks leading to interruption of communications, also has been observed a large number of blackouts. This data for optimization of civil infrastructure to avoid this crisis effects.

Big Data and Crisis Preparedness

Big Data analysis can help significantly to the preparation of crisis management. Through the data analysis can be done recognizing the dangers and to provide a sound strategic approach by the respective managers of the crisis. Big data analysis can also guide the proactive deployment of resources to fully cope with an impending type of disaster.

Big Data and Crisis Response

Big Data analysis in real time can identify which areas need the most urgent attention from the crisis administrators. With the use of the GIS and GPS systems, Big data analysis can assist the right guidance to the public to avoid or move away from the hazardous situation. Preceding crisis could help identify the most effective strategy for responding to future.

Big Data and Crisis Recovery

When the recovery activation will gradually start, the infrastructure would provide a big data source. The big data analysis sharing useful information for recovery procedures about volunteer coordination and logistics during the crisis.

Types of Crisis

In Crisis Administration process it label the different types of crises require the use of different crisis management strategies.

- Natural disaster,
- Technological crises,
- Confrontation,
- Malevolence, Organizational Misdeeds,
- Workplace Violence,
- Rumours,
- Terrorist attacks/man-made disasters.[13]

Crisis Administration information system is made up of a heterogeneous communication and computing network s address the crisis situation and adopt the network according to the service parameters. With a number of important corners and their corresponding constraints in taking network policies. To meet the unpredictable demands and accept various data formats, some subnets of the heterogeneous network is virtualized using virtualization. The heterogeneous network for crisis administration system is simulated as a mobile service.

An Crisis Administration system to deal with prevention and management of emergency situations which are either natural or man-made, the different system are designed so as to handle different calamities like flood, fire and terrorist attack etc. All the subsystems are managed, controlled and governed by government admin subsystem as either a centralized or a decentralized control.

The crisis administration application must handle a huge database which contains text, image, and maps; handle a huge audio video etc. A complex database management system must incorporated with the system .A large no queries are generated from the users during emergency situation .The queries that are raised by the different services and the users can handles by the physical networks. The Government authorities or the crisis administration with the help of the proposed computational model to analyse the situation Information in multiple formats is to be forwarded and processes in less time.[26]

The workflow management in emergency service computing environment that routes the key process activities and tasks to activate the emergency service based on the nature of emergency requests. The system considered, handle different types of request like messages, voice calls and sensor data from heterogeneous networks. The work flow model and its management make job simpler to take correct decisions before deploying the suitable type of service to the affected area. The model includes various sub-activities for continuous process improvements. The performances of collaborated activities in manual and automated services are monitored.

The performance depends on collaborated service level agreement whereby the requested service details have been mapped with or without negotiation. Service-level management is necessary to identify the potential problems and alert the society to minimize the risk of failure in mission-critical applications. Crisis administration is a multidisciplinary, multiorganizational, collaborative event requires organizing resources such as humans, technology, money, equipment. The work focus on the three research challenges.(a) designing a suitable workflow and coordination model for crisis events. (b) Designing a model for distributing workflow activities and (c) designing algorithms to perform recovery for workflow activities. The incoming emergency request may be received in form of short Messages called SMS message, telephone voice calls, E-mails and Data from sensor Networks .They will be classified and stored into the service queues after negotiating with requestors for coarse details of

the incidents. Processes are responsible in understanding the available services with the actual context of emergency request towards the expected volume of the service forces and monitoring the final deployment of the rescue units.

The service are carried via human activities like ultimate executive , service activation, management and monitoring The key performance indicator like the mean resource waiting time, cycle time, active time, maximum and mean resource cost for IP address checking, GPS network checking and service priority in the automated computing.[27]

In Smartphone number of emergency applications are available that are beneficial emergency response. Due to the innovations in GPS technology this can be very helpful in the tracking of assets and also helpful to send information to emergency management headquarters for the analysis of destruction caused by the disaster.[10]Three elements are common to a crisis Administration. a) Threat to the organization, b) the element of surprise and c) a short decision time

Conclusion

In the day to day life we handle the large amount of day sets that refers to as Bigdata.In this paper we review the recent bigdata 10 v's, bigdata technology, Hadoop, Mapreduce and crisis Administration .Crisis administration handle the emergency situation and overcome the situation .It is helpful to overcome emergency situation, preparedness and response.

References

1. Aaron Malveaux and A.Nicki Washington, A survey of Emergency Preparedness, International journal of Advanced computer Science and Applications vol. 6, No 7.2015
2. Adam Flizikowski, WitoldHolubowicz, Social Media in crisis Management the iSAR+Project survey.
3. Amir Gandomi, Murtazta Haider, Beyond the hype Bigdata concept, Methods, and Analysis., International journal of Information management, 137-144,35(2015).
4. Mr.AntonyPrakash,Dr.A.Aloysius, Survey on MapReduce on Big Data , International journal of Engineering and Computer Science ISSN:2319-7242 vol 6,Issues 3 pg:20662-20666, 3 March 2017
5. AnwaarAli,JunaidQadiret,, al , BigData For Development: Applications and Techniques arxiv:1602.07810v1 [cs.CY]25 Feb 2016.
6. Chaitalimohite et al, Survey Based on Hadoop, International journal for research in Applied Sccience and Engineering Technology(IJRASET) volume:5 Issue IV, ISSN 2321-9653,April 2017.
7. ChenXin,CUI Bin-ge et al,International Conference on Medical Physics and Biomedical Engineering,33(567- 572),2012.
8. Christian sell,Isris Braun, Using Workflow Management System to Manage Emergency Plans.
9. Dontas Emmonouil , Donkas Nikolas, Big Data Analytics in prevention, Preparedness, Response and Recoverry in Crisis and Disaster Management, Recent Advances in Computer science .
10. [10]HafsaMarjam, QaisarJavaid et al. A survey on Smartphone system for Emergency Management(SPSEM) , International journal of Advanced Computerscience and Application(IJACSA) vol 6, No 6, 2016.
11. https://www.google.co.in/search?q=hadoop+architecture&client
12. https://en.wikipedia.org/wiki/Crisis_management
13. http://en.wikipedia.org/wiki/Bigdata.
14. Jeffry Dean and Sanjay Ghemawet, MapReduce: Simplified Data Processing on Large Cluster, Google,Inc.

15. JyotiKumari, Mr.Surender,Statically Analysis on Big Data Using Hadoop , International journal of Computerscience and Mobile Computing,(IJCSMC) vol 6,Issue 6, pg 256-265, ISSN 230-088X
16. KuchipudiSravanthi, TatireddySubba Reddy, Applications of BigData in various fields, International journal of Computerscience and Information Technologies, vol 6,(5), 4629-4632 ,ISSN:0975-9646, 2015.
17. Manish Kumar Khani, SweetyKakhani and S.R Biradar, Research Issuess in Big Data Analytics, volume 2, Issues 8 ,ISSN 2319-4847, August 2013.
18. A.Mobaraki, A.Mansourian,etal,Application of Mobile GIS and SDI for Emergency Management.
19. NawSherrkhan,IbrarYaqoob et al, Big Data: Survey, Technologies, Opportunities, and challenges, Hindawi
20. Publishing Corporation,The Scientific World Journal, volume 2014,Article IDb712826,pg
21. http://dx.doi.org/10.1155/2014/4/712826.
22. PriyankaRana, A Study on BigData:Concept, Architecture, Technologies,ISSN 2320-9798, vol 5, Issue 4, April 2017.
23. Ms.Rucha S Shelodkar ,Prof.Himanshu U Joshi, On The Fly MapReduce Aggregation for Big Data Processing In Hadoop Environment, International Journal of Computational Engineering Research(IJCER), volume 07, Issue 07,ISSN(e)2250-3005,july 2017.
24. Samithasahu et al, Research Issues and Challenges of Bigdata –A Review, International Journal of Scientific Engineering and Applied science(IJSEAS) volume-2 issue-5 May 2016 ISSN :2395-3476.
25. G.Suganya MCA, A Study of Big Data Mining Techniques Used to Improve the Performance of Large Data Processing, International Journal ofInnovative Research in Computer and Communication Engineering, vol 5,Issue 7, July 2017
26. Varsha B.Bobade."Survey paper on BigData and Hadoop" International research Journal of Engineering and Technology (IRJET) ISSN :2395-0056 vol 03, Issue 01,pp 861.

BIG DATA WITH INTERNET OF THINGS (IOT): A REVIEW

R.Madhubala

M.Phil. Scholar, Department of Computer Science, Raja Doraisingam Government Arts College, Sivagangai

A.Prema & K.Chelladurai

Assistant Professors, Department of Computer Science, Raja Doraisingam Govt Arts College, Sivagangai

Abstract

The Internet of things (IOT) and big data go together. The main idea behind the IOT is that almost every object or device will have an IP address and will be linked with each other Considering the fact that trillions of devices will be linked and will be producing massive volumes of data, the efficiency of data collection mechanism is going to be challenged. Internet of things (IOT) is a promising paradigm in the integration of several technologies and communication solution. Analyzing big data is a challenging task as it involves large distributed file system which should be fault tolerant, flexible and scalable. Hadoop is used for processing massive data which use map reduce programming model. This survey is on the discussion of open problems and future direction of integration big data and internet of things (IOT).

Keywords: *Internet of Things (IOT); Big Data; Hadoop; Map reduce.*

Introduction

Big data information comes from various, heterogeneous, autonomous sources with complex relationship continuously growing. Upto 2.5 quintillion bytes of data are created daily and 90 percent data in the world today were produced within past two years [15]. For example Flicker, a public picture sharing site where in an average of 1.8 million photos per day have been received from February to march 2012[5].This shows that it is very difficult for big data applications to manage, process and retrieve data from large volume of data using existing software tools. It's become achallenge to extract knowledgeable information for future use. Nowadays, big data related to the service of Internet companies grow rapidly. For example, Google processes data of hundredsof Petabyte (PB), Facebook generates log data of over 10 PB per month, Baidu, a Chinese company, processes data of tens of PB, and Taobao, a subsidiary of Alibaba, generates data of tens of Terabyte (TB) for online trading per day [6].

Internet of Things

The Internet of Things (IoT) is the original paradigm that is rapidly gaining ground in the scenario of modern wireless telecommunications. The basic idea of this concept is the pervasive presence of a variety of things or objects around us such as Radio-Frequency Identification (RFID) tags, sensors, actuators, mobile phones, etc. – which, through unique addressing schemes, are able to interact with each other and cooperate with their neighbours to reach common goals [4].The very first definition of IoT derives from a ''Things oriented" perspective; the considered things were very simple items: Radio-Frequency Identifications (RFID) tags. The terms ''Internet of Things**s**" is, in fact, attributed to The Auto-ID Labs [1].

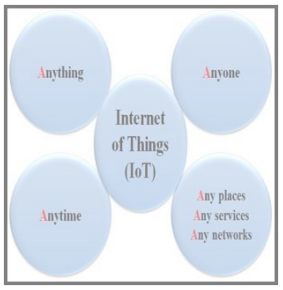


Figure 1 Internet of Things

Internet of things, anything's will able to communicate to the internet at any time from any place to provide any services by any network to anyone. this concept will create a new types of applications can involve such as smart vehicle and the smart home, to provide many services such as notifications, security, energy saving, automation, communication, computers and entertainment.

In the IoT, the communication language will be based on interoperable protocols, operating in heterogeneous environments and platforms [8]. IoT in this situation is a basic term and all objects can play an active role to their connection to the Internet by creating smart environments, where the role of the Internet has changed

Big Data

Big data is an intellectual concept. Apart from masses of data, it also has some other features which decide the difference between itself and "**massive data**" or "**very big data**."The need of big data generated from the large companies like Facebook, Yahoo, Google, YouTube etc.

Google contains the large amount of information. So there is the need of Big Data Analytics that is the processing of the complex and massive datasets and this data is different from structured data in terms of ten parameters –variety, volume, value, veracity, velocity, validity, variability, venue, vocabulary and vagueness (10V's) [3].

The challenges of big data management are

1. **Volume:** Data is ever-growing day by day of all types MB, PB, YB, ZB, KB, TB of information. The data results into large files. Excessive volume of data is the main issue of storage. This main issue is resolved by reducing storage cost.Data volumes are expected to grow 50 times by 2020.
2. **Variety:** Data sources are extremely heterogeneous. The file comes in various formats and of any type, it may be structured or unstructured such as text, audio, videos, log files and more. The varieties are endless, and the data enters the network without having been quantified or qualified in any way.
3. **Velocity:** The data comes at high speed.Sometimes 1 minute is too late, so big data is time sensitive. Some organisations data velocity is a main challenge. The social media messages and credit cardtransactions are done in millisecond and data generated by this putting in to databases.
4. **Value:** It is the most important v in big data. Value is main buzz for big data because it is important for businesses, IT infrastructure system to store large amount of values in database.
5. **Veracity:** The increase in the range of valuestypical of a large data set. When we deal with high volume, velocity and variety of data, the all of data are not going 100% correct, there will be dirty data. Big data and analytics technologies work with these types of data.
6. **Validity:** Similar to veracity, validity refers to how accurate and correct the data is for its intended use. The benefit from big data analytics is only as good as its underlying data, so you need to adopt good data governance practices to ensure consistent data quality, common definitions, and metadata.
7. **Variability:** Variability in big data's context refers to a few different things. One is the number of inconsistencies in the data. These need to be found by anomaly and outlier detection methods in order for any meaningful analytics to occur. Variability can also refer to the inconsistent speed at which big data is loaded into your database.
8. **Vocabulary:** Data science provides a vocabulary for addressing a variety of problems. Different modeling approaches tackle different problem domains and different validation techniques harden these approaches in different applications.
9. **Venue:** Distributed heterogeneous data from multiple platforms, from different owners systems, with different access and formatting requirements, private vs. public cloud.
10. **Vagueness:** The meaning of found data is often very unclear, regardless of how *much* data is available.

Relations between IoT and big data

The big data generated by IoT has different characteristics compared to general big data because of the different types of data collected, of which the most classical characteristics include heterogeneity, variety, unstructured feature, noise, and high redundancy. A report from Intel pointed out that big data in IoT has three types that conform to the big data paradigm:

- Abundant terminals generating masses of data;
- Data generated by IoT is usually semi-structured or unstructured;
- Data of IoT is useful only when it is analysed;

There is a compelling need to accept big data for IoT applications, while the development of big data is already legged behind. It has been widely recognized that these two technologies are inter-dependent and should be jointly developed: on one hand, the widespread deployment of IoT drives the high growth of data both in quantity and category, thus providing the opportunity for the application and development of big data; on the other hand, the application of big data technology to IoT also accelerates the research advances and business models of IoT.

Application of IoT based big data

IoT is not only an important resource of big data but also one of the main markets of big data applications. Because of the high variety of objects, the applications of IoT also evolve endlessly. Logistic enterprises may have profoundly experienced with the application of IoT big data. For example, trucks of UPS are equipped with sensors, wireless adapters, and GPS, so the Headquarter can track truck positions and prevent engine failures. Meanwhile, this system also helps UPS tosupervise andmanage its employees and optimize delivery routes. The optimal delivery routes specified for UPS trucks are derived from their past driving experience. In 2011, UPS drivers have driven for nearly 48.28 million km less [11].

HADOOP

Hadoop

Hadoop is a structure that can run applications on frameworks with a large number of hubs and terabytes. Hadoop architecture shown in Fig 2 It distributes the file among the nodes and allows the system to continue work in case of a node failure. This approach reduces the risk of data strophic system failure [8].

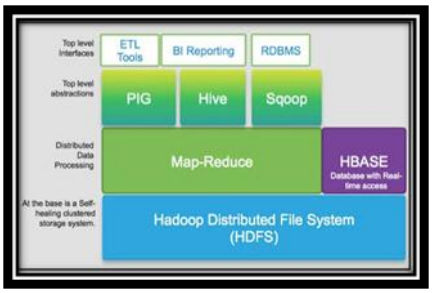


Figure 2 Hadoop Architecture

Components of Hadoop

HBase: It is an open source circulated and Non-socialdatabase framework executed in Java. It runs above the layer of HDFS. It can serve the input and output for theMap Reduce in well-mannered structure.

Oozie: Oozie is a web-application that runs in a java\servlet. Oozie uses the database to gather the information of Workflow which is a collection of actions.

Sqoop: Sqoop is an order line interface application that gives stage which is accustomed to changing over datafrom relational databases and Hadoop or vice versa.

Avro: It is a framework that gives usefulness of information serialization and administration of information trade. It is basically used in Apache Hadoop. These services can be used together as well as independently according the data records.

Chukwa: Chukwa is a structure that is utilized for information gathering and investigation to handle andbreak down the massive amount of logs. It is built on the upper layer of the HDFS and Map Reduce framework.

Pig: Pig is high-level platform where the Map Reduce frame work is created and used with Hadoop platform. It is a high level data processing system where the data records are analyzed occuring in high level language.

Zookeeper: It is centralization based administration that gives conveyed synchronization and gives group services along with maintenance of the configuration information and records.

Hive: It is application developed for data warehouse that provides the SQL interface as well as relational model. Hive communications is built on the top layer of Hadoop that help in provides conclusion and analysis for respective queries.

Applications of Hadoop

- Few of the applications of Hadoop are given below [12][14] [2]
- Log and/or clickstream analysis of various kinds
- Marketing analytics
- Online travel booking
- Energy discovery and energy savings
- Infrastructure management
- Fraud detection
- Health care

Tracks various types of data such as Geo-location data, machine and sensor data, social media data

Map Reduce

Map Reduce is a programming model created by **Google**. Map reduce is an encoding model designed for dispensation volumes of data in parallel by dividing the work into set of independenttasks.It has been designed for simplifying parallel data processing on large clusters. Before developing the Map Reduce Framework, Google used hundreds of different implementation to process and compute large dataset. Most of the input data was very large but the computations were relatively simple. Hence the computations needed to be scattered across hundreds of computers in order to finish calculations in a reasonable time. [9]

Programming Model of Map Reduce

The nature of this programming model and how it can be used to write programs which run in the Hadoop environment is explain by this model. Hadoop is an open source implementation for this environment [10]. Map Reduce programming model consists of data processing functions: ***Map*** and ***Reduce***. Parallel Map tasks are run on input data which is partitioned into fixed sized blocks and produce intermediate output as a collection of <key, value> pairs. These pairs are shuffled across different reduce tasks based on <key, value> pairs. Each Reduce task accepts only one key at a time and process data for that key and outputs the results as <key, value>pairs [7].

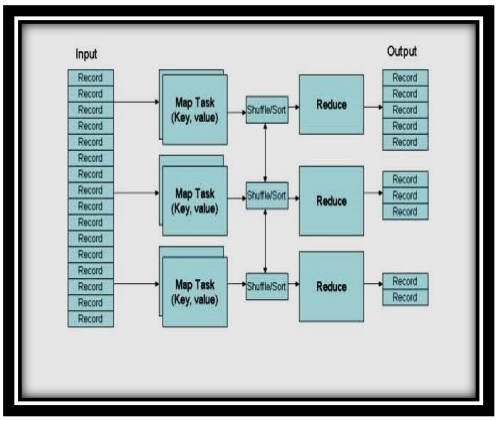


Figure 3 Programming Model of Map Reduce

Application of Map Reduce

Map Reduce can be used in SMS message mining. SMS messages are well-liked and broadly used for simple communication among people. Number of SMS messages sent in a month in any country is very large, and so is the original dataset used mining. Finding the most popular SMS messages can be precious information but since the dataset is so large, parallelization is needed to complete this task in realistic time. Hadoop, the open-source implementation of Map Reduce, is used as a framework in SMS mining. Processing of the message is done in three steps, [13]

First the original dataset is pre-processed and grouped by sender' mobile numbers. This is done by first Map Reduce process

Second Map Reduce process does an alteration to regroup the dataset by short content keys.

Third Map Reduce phase is needed to mine the popular messages.

Conclusion

Big data will not only have the social and economic impact but also influence everyone's waysof living and thinking which is just happening. We could not predict the future but may take precautions for possible events to occur in the future. With Big Data technologies we are capable to provide most relevant and accurate social sensing feedback to better understand society at real-time. We believe that given the interest shown by industries in the IoT applications, in the next years addressing such issues will be a powerful driving factor for networking and communication research in both industrial and academic laboratories.

References

1. <http://www.autoidlabs.org/>. Business Applications of Hadoop.
2. http://www.edureka.in/blog/business-aplications-of-hadoop/
3. Chris Eaton, Dirk Deroos, Tom Deutsch, George Lapis, and Paul Zikopoulos, Understanding Big Data: McGraw-Hill Companies, April 2012,
4. D. Giusto, A. Iera, G. Morabito, L. Atzori (Eds.), The Internet of Things, Springer, 2010. ISBN: 978-1-4419-1673-0.
5. F.Michel, "How Many Photos Are Uploaded to Flickr Every Day and Month?"http://www.flickr.com/photos/franckmichel/6855169886/, 2012.
6. Hadoop. http://hadoop.apache.org/.
7. International Journal of Computer Applications *(0975 – 8887) "*Survey on Improved Scheduling in Hadoop Map Reduce in Cloud Environments "Volume *34– No.9, November 2011.*
8. Jimmy Lin "Map Reduce Is Good Enough?" The control project. IEEE Computer 32 (2013)
9. J. Dean and S. Ghemawat. Map reduce: simplified data processing on large clusters. Commun. ACM, 51(1):107–113, 2008
10. K. V. Shvachko and A.C. Murthy, "Scaling Hadoop to 4000 Nodes at Yahoo" Yahoo! Developer Network Blog, 2008
11. Min Chen, Shiwen Mao, Yunhao Liu "Big Data: A Survey" Mobile Netw Appl (2014) 19:171– 209 DOI 10.1007/s11036-013-0489-0
12. https://en.wikipedia.org/wiki/ApacheHadoop
13. T. Xia. Large-scale sms messages mining based on map-reduce. Computational Intelligence and Design, 2008. ISCID '08.International Symposium on, 1:7–12, Oct. 2008
14. https://gigaom.com/2012/06/05/10ways-companies-are-usinghadoop-to-do-more-than-serve-ads/
15. Xindong Wu, Fellow, IEEE, Xingquan Zhu, Gong-Qing Wu, and Wei Ding" Big Data with data mining" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 1, JANUARY 2014
16. Image: Utilizing hadoop &HDFS as an active archiving &storage
17. Image: Map reduce programming model problems…https://goo.gl/images/SKaquj.
18. Internet of Things Applications, www.researchgate.net

TOOLS AND TECHNOLOGIES USED IN BIG DATA AND HADOOP A REVIEW

K.Revathi

M.Phil. Scholar, Department of Computer Science, Raja Doraisingam Government Arts College, Sivaganga

A.Prema

Assistant Professor, Department of Computer Science, Rajadoraisingam Government Arts College, Sivaganga

Abstract

We are living in an enormously developed, technical era, where internet is fundamental need. Today our professional life is revolving around World Wide Web giving birth to big data. The term 'Big Data' refers to datasets whose size, complexity and velocity make them difficult to capture, manage and analyze. Hadoop is the core platform for structuring big data and solved problem of making it for analytics purpose. In Hadoop a program known as Map-Reduce is used to collect data according to query. This paper focuses on various techniques used in big data and Hadoop Map-Reduce.

Keywords: *Big Data, Hadoop, Map-Reduce.*

Introduction

Imagine a world without data storage, a place where every detail about a person or organization, every transaction performed or every aspect which can be documented is lost directly after use. Data is the building block upon which any organization thrives. Now think of the extend of details and the surge of data information provided nowadays through the advancements in technologies and the internet. With the increase in storage capabilities and methods of data collection huge amount of data have become easily available. There have been 6 billion mobile subscriptions in the world and everyday 15 million text message are sent. By the year 2020, 50 billion devices will be connected to networks and the internet[10].

Big Data and Hadoop

Benefits of Big Data

When Big Data is meritoriously captured, handled and analyzed , companies are able to increase a more complete accepting of their business, customers, products competitors etc,which can lead to efficient improvements , increased sales ,lower costs ,better customer service and improved products and services.

Using information Technology (IT) logs to improve IT troubleshooting and security breach detection, speed, effectivesness, and future occurrence prevention Use of financial market transaction information to more quickly assess risk and take corrective action.

12v's Characteristics of Big Data

Volume

Volume is probably the best known characteristics of Big Data. Considering more than 90 percent of all today's data was created in the past couple of years. The current amount of data can actually be quite staggering.

Velocity

Velocity refers to the speed at which data is being generated, produced, created or refreshed.

Variety

Today data comes in different type of formats. Structured and numeric data in traditional databases Information created from line of business applications. Data variety exploded from structured and legacy data stored in enterprise storages to unstructured, semi structured, audio , video etc..

Variability

Variability in Big Data's context refers to few different things. one is the number of inconsistencies in the data. Variability can also refer to the inconsistent speed at which Big Data is loaded into our database.

Figure 1 Big Data

Veracity

This is one of the unfortunate characteristics of Big Data. Veracity refers more to the derivation or reliability of the data source, itsframework , and how meaningful it is to the analysis based on it.

Validity

Big Data veracity is a matter of validity, meaning that the data is correct and accurate for the intended use. Clearly valid data is the key for making the right decisions[9]. Data validation is one that certifies uncorrupted transmission of data.

Volatility

How old does our data need to be before it is considered irrelevant , historic, or not useful any longer? How long does data need to be kept for? Before big data establishmentsinclined to store data indeterminately a few terabytes of data might not create high storage expenditures. It could even be kept in the live database without causing performance issues.

Visualization

Another characteristics of big data is how challenging it is to visualize. Present big data visualization tools face methodologicaltests due to limitations of in-memory technology and poor scalability, functionality, and response time.

Value

Last , but arguably the most important of all , is Value. Substantial value can be found in big data , including understanding your customers better, targeting them accordingly , optimizing process , and improving machines or business performance[9].

Verbosity

Verbosity particularly relates to questions about text sources and the problems of machine understanding of the meaning of the text.

Vulnerability

Vulnerability poses a wide range of questions about the system, the data and the stakeholders.

Verification

It help us to demonstrate that we understand and have addressed the vulnerabilities [16].

Tools of Big Data

Python: Python is a potent , malleable , open – source language that is easy to learn , easy to use , and use powerful libraries for data operation and analysis.

R: R is an open source programming language and software environment for statistical computing and graphics.

Hadoop: The name Hadoop has become synonymous with big data. Hadoop is an open-source software structure for scattered storage of verylarge datasets on computer clusters. Hadoop provides massive amounts of storage for any kind of data, enormous processing power and the ability to handle virtually limitless concurrent tasks or jobs. Hadoop is not for the data beginner. It distributes the file among the nodes and allows to system continue work, in case of a node failure. This approach reduces the risk of catastrophic system failure. In which application is broken into smaller parts (fragments or blocks). Apache Hadoop consists of the HadoopKernal, Hadoop distributed file system (HDFS), Map-Reduce and related projects are Zoo keeper , Apache Hive. Hadoop Distributed File System consists of three components: The Name node, Secondary Name Node and Data Node.

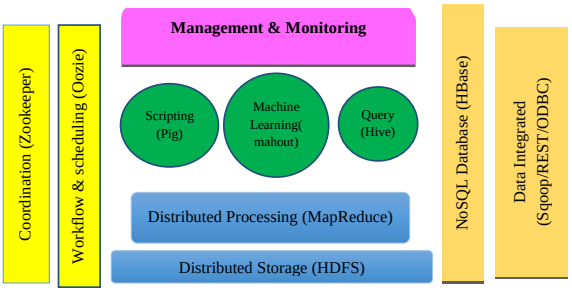


Figure 2: Apache Hadoop Ecosystem[2]

- **Hive:** Hive is a "SQL-Like" connection that allows predictable BI applications to run queries against a Hadoop cluster.
- **PIG:** PIG is another connection that tries to conveyHadoop closer to the actualities of developers and business users, similar to Hive.
- **WibiData:** Wibi Data is a combination of web analytics with Hadoop being built on top of HBase , which is itself a database layer on top of Hadoop.
- **Platfora:** PLATFORA is a platform that turns user's queries into Hadoop jobs spontaneously , thus generating an abstraction layer that anyone can adventure to simplify and establish datasets stored in Hadoop.
- **Sky Tree:** Sky Tree is a high-performance machine learning and data analytics platform focused specifically on handling Big Data.
- **Tool:** Data visualization is a modern branch of descriptive statistics. It involves the creation and study of the visual representation of data. Which means " Informationthat has been abstracted in some schematic form , including attributes (or) variables for the units of information". Some of the tools are, D3 , Data Wrapper , Tableau[11].
- **Map Reduce:** Map-Reduce is a similar programming model for writing scattered applications developed at Google for effective processing of large amounts of data, on large clusters of commodity hardware in a reliable , fault – tolerant manner. The Map-Reduce program runs on Hadoop which is an Apache open-source frame work. The Map-Reduce algorithm comprises two important tasks, namely Map and Reduce. Map remains a set of data and converts it into another set of data, where separate elements are broken down into tuples. Secondly condensetasks , which takes the output from a Map as an input and syndicates those data tuples into a smaller set of tuples. The major benefit of Map Reduce is that it is easy to scale data handling over multiple computing nodes[4].

	Apache Hadoop	IBM Info Sphere
Mode of software	Open source and free source	Commercial
Type of Data	Unstructured data, time series, textual data	Unstructured data, semi structured data and structured data.
Data Sources	Files, the network scripted Output	IBM Warehouse
Database Support	HBASE, Sybase, SAP	Mongo DB , DB2 , Oracle
Operating System	Windows, Linux	Windows

Table 1comparison of Big Data Tools [10]

Big Data Tools Techniques Tools and Technologies

To capture the value from Big Data, we need to develop new techniques and technologies for analyzing it. Until now scientist have developed a wide variety of techniques and technologies to apture, curate, analyze, and visualize Big Data. These techniques and technologies cross a number of discipline, including computer science, economics, mathematics, statistics, and other expertises. Multidisciplinary methods are needed to discovery the valuable information from Big Data. We need tools make sense of Big Data. Current tools concentrate on three classes, namely, batch processing tools, stream processing tools, interactive analysis tools. Most batch processing tools are based on the Apache Hadoop infrastructure, such as Mahout and Dryad. The latter is more like necessary for real-time analytic for stream data applications. Storm and S4 are good examples for large scale streaming data analytic platforms.

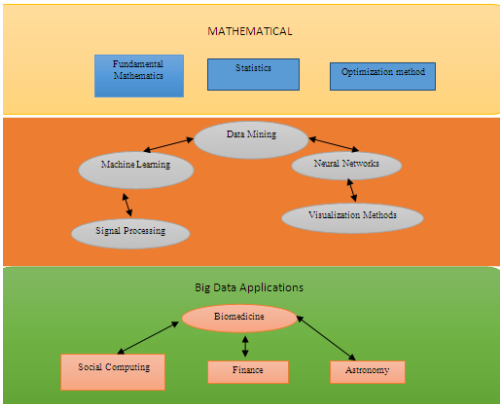


Figure 3: Big Data Applications

Big Data Techniques

Big Data needs extraordinary techniques to efficiently process large volume of data within limited run times. Reasonably, Big Data techniques are driven by specified applications. For example, Wal-Mart applies machine learning and statistical techniques to explore patterns from their large volume of transaction data. Big Data techniques involve a number of disciplines , including statistics , data mining , machine learning , neural networks ,social network analysis , signal processing , pattern recognition , optimization methods and visualization approaches[17].

Big Data Tools Based on Batch Processing

One of the most important and powerful batch process –based Big Data tools is Apache Hadoop. It provides infrastructures and platform for other specific Big Data applications. A number of specified Big Data systems are build on Hadoop, and have special usages in different domains ,for example Data mining and Machine learning used in business and commerce.

Stream Processing Big Data tools

Hadoop does well in processing large amount of data in parallel. It provides a general partitioning mechanism to distribute aggregation workload across different machines. Nevertheless, Hadoop is

designed for batch processing. It is a multi-purpose engine but not a real-time and high performance engine, since there are high throughout latency in its implementations.

Big Data tools based on interactive analysis

In recent years, open source Big Data systems have emerged to address the need not only for scalable batch processing and stream processing, but also interactive analysis processing .The interactive analysis presents the data in an interactive environment, allowing users to undertake their own analysis of information.

C.Lakshmietal defined the term big data. Big data is a largest buzz pharses in domain of IT. New technologies of personal communication driving the big data new trend and internet population grew day by day but it never reach by 100%. The need of big data generated from the large companies like Facebook, yahoo, Google, you tube etc. for the purpose of analysis of enormous amount of data which is an unstructured form or even in structured form[5][2].C.Lakshmi, V.V.Nagendra Kumar etal described the technologies that are used in big data, and explained about the component of Hadoop[5]. The (HDFS) Hadoop distributed file system is the file system component of the Hadoop frame work. HDFS was designed and optimized to store data over a large amount of low-cost hardware in a distributed fashion.

Mr.PiyushBhardwaj etal defined 5v's of big data and described the big data tools [10].Due to big data, new bets are being formed by Netflix. Netflix analyses 30 million "plays" a day, including when it is paused ,rewind and fast forward, four million ratings by Netflix subscribers, three million searches as well as the time of day when shows are watched and the extensive user account information. The show "house of cards" was a result of analysis of data and generated a profit of millions of dollars with the lead and direction as Kevin Spacey and David fincher respectively, eventually resulted a big hit. the main problem that we face today is how to convert unstructured data to structured data. AbhishekGupta ,Kataletal described the term unstructured data [8].It refers the information which was scattered and not organized in a proper format. Unstructured data is human material similar that emails, videos, tweets, facebook posts, request chats, closed circuit Tv footage, mobile phone calls, website clicks. Unstructured data is considered as " loosely structured data" because the data sources possesses a structure but all the data within a dataset donot have a particular.

Mrs.Mereena Thomas et al explained 5v's of Big Data (volume, velocity, veracity, variety and value) [7].Volume refers to the enormous amount of data which is produced and processed to meet demand. Veracity refers to the messiness of the data. Variety refers to the different types of data we can now use. Value refers to our ability turn our data into value[7].

Harshawardhan S.Bhosale & Prof.Devendra P.Gadekar et al were illustrated the Hadoop Map Reduce[4] Hadoop creates cluters of machines and co-ordinates work among them, clusters can be built with inexpensive computers. If one fails, Hadoop continues to operate the cluter without losing data by shifting work to the remaining machines in the clusters[8]. Hadoop is a programming framework used to support the processing of huge data sets in a scattered computing environment.

MikinK.Dagli et al describes the concept of Big data and Hadoop implementation for processing and generating large data sets with a parallel, distributed algorithm on a cluter[1].MapReduce programming model has been successfully used at Google for many different purposes, success of Mapreduce is based on various reasons[3].

YashikaVerma et al, illustrated the concept of Big Data along with 5V's. It also focuses on Big Data processing problems. These technical challenges must be addressed for efficient and fast processing of Big Data [15]. They described the data storage techniques. It used for Big Data includes multiple clustered

network attached storage and object based storage.[15]. Big Data analytics is where analytic techniques are applied on big data sets. Such data sets can no longer be easily analyzed with traditional data management and analysis techniques and infrastructures. Analytics based on large data samples reveals and leverages business change.[11]They are also defined the three main features of Big Data.

Nada SaelHussein. SuhaliOwais, Zhang L. Stoffel et al describes the Visual analytics for the Big Data and 9V's of Big Data[9].

The framework maps the different big data storage, management and processing tools, analytics tools and methods and visualization and evaluation tools to the different phases of the decision making process. Hence the changes associated with the big data analytics are reflected in three main areas: big data storage and architecture, data and analytics processing, and, finally, and the big data analyses which can be applied for knowledge discovery and informed decision making.

Conclusion

Big Data is going to more diverse, larger, faster and is becoming the new scientific data research and for business applications. In this paper we have surveyed to handle the Big Data and also discussed tools that are used in Big Data and Hadoop. A success of MapReduce is based on various reasons. First the model is easy to use. Second a large set of problems with various nature are easily expressible as MapReduce computations. Second a large sets of problems with various nature are easily expressible as MapReduce computations. Third, MapReduce can implement on large clusters of commodity hardware. The main goal of our paper was to make a survey of various big data handling techniques those handle a massive amount of data from different sources and overall performance of systems.

References

1. AbdulRaheem Syed, Kumar Gillela, Dr.C.Venugopal, "The Future Revolution on Big Data", In International Journal of Advanced Research in Computer and Communication Engineering, e-ISSN: 2278- 1021, p-ISSN: 2319-5940, Volume:2, Issue:06, P.No:2446-2451,2013.
2. ChennupatiVinay Kumar, "Survey on Big Data Analytics and Hadoop Tools", International Journal of Science Technology and Engineering, Volume: 3, Issue: 07, ISSN: 2349-784X, January 2017.
3. MikinK.Dagli, BrijeshB.Mehta, "Big Data and Hadoop: A Review "In IJARES , ISSN:2347-9337, Volume: 2, Issue: 2, P.No:192, Feb-2014.
4. HarshawardhanS.Bhosale Prof.DevendraP.Gadekar, "A Review Paper on Big Data and Hadoop", International journal of scientific and Research Publications, Volume: 4, Issue: 10, October-2014.
5. C.Lakshmi, V.V.Nagendra Kumar, "Big Data", International journal of Advanced Research in Computer Science and Software Engineering, ISSN:2277-128X, Volume:06, Issue:08, P.No:368-381, August-2016.
6. ManjulaDyavanur, KavitaKori, "Fault Tolerance Techniques in Big Data Tools: A Survey", International journal of Innovative Research In Computer and Communication Engineering, Volume: 02, issue: 02, ISSN: 2320-9801, ISSN: 2320-9798, May-2014.
7. Mrs.Mereena Thomas, "A Review Paper On Big Data", International Research Journal of Engineering and Technology (IRJET), e-ISSN: 2395-0056, p-ISSN: 2395-0072, Volume: 02, Issue: 09, Dec-2015.
8. Mohammad Wazid, Katal, RH.Goudar, "Big Data: Issues, Challenges, Tools and Good Practices, IEEE 978-1-4799-0192-0/28, May 2014.

9. Nada SaelHussein, SuhaliOwais, "Extract five Categories Cpivw from the 9V's Characteristics of the Big Data", International journal of Advanced Computer science and application, Volume: 7, Issue: 03, 2016.
10. PiyushBhardwaj, "Comparative Analysis of Big Data Tools", International journal of Computer science and Mobile Computing, Volume: 05, Issue: 05, P.No:789-793, May-2016.
11. Rama Devi Gunnam, "Importance and Tools used: Big Data", International journal of Advanced Research in Computer science and Software Engineering, Volume: 06, Issue: 05, P.No:513-516, May-2016
12. SumanArora, Dr.MadhuGoel, "Survey Paper on Scheduling in Hadoop", International journal of Advanced Research in Computer Science and Software Engineering, Volume: 04, Issue: 05, May – 2014.
13. VeenakshiDevi, Meenakshi Sharma, "Survey on Big Data Tools and Techniques", ISOR Journal of Computer Engineering (ISOR-JCE), e-ISSN: 2278-0661, p-ISSN: 2278-8727, 2014.
14. Mrs.VibhavariChavan, Prof.Rajesh, N.Phursule, "Survey paper on Big Data", International journal of Computer Science and Information Technologies, Volume: 05, Issue: 06, 2014.
15. YashikaVerma ,SumitHooda , "A Review Paper on Big Data and Hadoop", International Journal for Scientific Research and Development, Volume: 03,, Issue : 02, P.No:170 , 2015.
16. Richard J Self," The 12 Vs of Big Data Governance", Research Fellow, Big Data Laboratory University of Derby.
17. C.L .Philip chen, and Chun-Yang Zhang, Data-intensive applications, challenges, techniques and Technologies: A survey on Big data, Elsevier 2014.
18. M.S VibhVarichavan, Prof Rajesh.N.Phursule, Survey paper on BigData, International journal of computer Science and Information Technologies vol 5, Issue 6, ISSN:7932-7939,2014

A STUDY ON WEB USAGE MINING FOR WEB PERSONALIZATION

M.Muthalagu

Assistant Professor, PG Department of Computer Science, Thiagarajar College, Madurai

Abstract

Web is an enormous store house of Information. Web Mining is an upcoming trend which serves as per the user requirements. It has many in types. Web Usage mining is to discover interesting user navigation patterns and can be applied to many real world problems, such as improving website/pages, making additional customer behavior studies etc. Personalization is a subclass of information filtering system that seeks to predict the ratings or preferences that a user would give to an item, they have not yet considered using a model built from the characteristics of the item. This proposed system uses information such as profile info, time being spent by the user, preferences from log file then process the selected info and then improvise the users recommendation as per the results.

Keywords: *Web mining, Web usage mining, Customization, Web personalization*

Introduction

Data mining refers to extracting or "**mining**" knowledge from large amounts of data. Knowledge Discovery in Data is the non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data. Data mining consists of more than collection and managing data; it also includes analysis and prediction. People are often do mistakes while analyzing or, possibly, when trying to establish relationships between multiple features.

Web Mining

Web mining techniques are commonly used by various kinds of organization to extract useful information on the basis of interests and habits of consumers and users. The extracted information used in business, market dynamics, new promotions on internet, personalization advertisements etc. [7].

Web Usage Mining

Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data in order to understand and better serve the needs of Web-based applications. Usage data captures the identity or origin of Web users along with their browsing behavior at a Web site. Web Usage Mining is to mine data from log record on web page. Log records lots useful information such as URL, IP address and time and so on. Analyzing and discovering Log could help us to find more potential customers and trace service quality and so on. The web usage mining is the process of applying the data mining technology to the web data and is the pattern of extracting something that the users are interested in from their network behaviors.

When people visit one website, they leave some data such as IP address, visiting pages, visiting time and so on, web usage mining will collect, analyze and process the log and recording data. Through these, utilize some mathematic method to establish users' behaviour and the interest models, and use these models to understand the user behaviour, thus to improve the website structure. Then finally provides a better characteristic information service for the user. The content and structure of a Web site are used as inputs to every major step of the process. Web usage mining itself can be classified further depending on the kind of usage data considered.

- **Web Server Data**: The user logs are collected by the Web server. Typical data includes IP address, page reference and access time.
- **Application Server Data**: Commercial application servers have significant features to enable E-commerce applications to be built on top of them with little effort. A key feature is the ability to track various kinds of business events and log them in application server logs.
- **Application Level Data**: New kinds of events can be defined in an application, and logging can be turned on for them thus generating histories of these specially defined events.

Personalization

Personalization means persons would get the things or results according to their interests and expectations without giving much more input. Users want the Wide World should shrink into a tab as he/she requires way. So there should be a system that analysis the user's expectation & preferences explicitly from user profiles or implicitly from user log files. These types of systems are developed from the concept of Web Mining. So these systems personalize a web site as per the user needs [6].

Literature Review

Pooja Mehta. (2012) has described as Web mining is the application of data mining techniques to extract knowledge from Web data including web documents, hyperlinks between documents, usage logs of web sites, etc.[6]. Data may consist of text, images, audio, video, or structured records such as lists and tables.

A.J Ratnakumar (2005) has described as Web mining has 3 more types of interesting research areas. They are i) Web Content Mining ii) Web Structure Mining iii) Web usage mining. Web Content Mining is Web Content Mining is the process of extracting useful information from the contents of Web documents. Content data corresponds to the collection of facts a Web page was designed to convey to the users. It may consist of text, images, audio, video, or structured records such as lists and tables [3].

R. Girardi (2007) has described as Web usage mining aims at discovering interesting patterns of use by analyzing Web usage data. It is the process of applying data mining techniques to the discovery of usage patterns from Web data generated by user interactions with a Web server, including Web logs, click streams and database transactions at a Web site or at a group of related sites [4].

M. Eirinaki (2005) has described as The objective of a Web personalization system is to provide users with the information they want or need, without expecting from them to ask for it explicitly [2].

W. Bin *et al.* (2003) has described as Personalization requires implicitly or explicitly collecting visitor information and leveraging that knowledge in your content delivery framework to manipulate what information you present to your users and how you present it. A personalization mechanism is based on explicit preference declarations by the user and on an iterative process of monitoring the user navigation, collecting its requests of ontological objects and storing them in its profile in order to deliver personalized content [1].

A.J Ratnakumar (2005) has described as The Web Personalization process divides in to four distinct phases [3].

- **Collection of Web Data**–In this, implicit data includes past activities/click streams as recorded in Web server logs and/or via cookies or session tracking modules. Explicit data usually comes from registration forms and rating questionnaires.

- **Preprocessing of Web Data**–In this, Data is frequently pre-processed to put it into a format that is compatible with the analysis technique to be used in the next step. Preprocessing may include cleaning data of inconsistencies, filtering out irrelevant information according to the goal of analysis
- **Analysis of Web Data**- This step applies machine learning or data mining techniques to discover interesting usage pattern and statistical correlation between web pages and user groups. This step frequently results in automatic user profiling, and is typically applied offline, so that it does not add a burden on the web server.
- **Decision making/Final Recommendation-** It makes use of the results of the previous analysis step to deliver recommendations to the user. It involves generating dynamic Web content on the fly, such as adding hyperlinks to the last web page requested by the user.

M. Aarti et al (2015) Analyzing the web log files through web usage mining is very important to discover the similar behavior users of particular website. Our paper discusses how to find useful knowledge from web log file using some data mining technique like Association rule mining and clustering. First we preprocess the web log file then apply association rule mining and clustering algorithm on web log file to discover usage pattern and same behavioral users.

Suresh shirgave et al (2013) has described Explosive and quick growth of the World Wide Web has resulted in intricate Web sites, demanding enhanced user skills and sophisticated tools to help the Web user to find the desired information. Finding desired information on the Web has become a critical ingredient of everyday personal, educational, and business life. Thus, there is a demand for more sophisticated tools to help the user to navigate a Web site and find the desired information. The users must be provided with information and services specific to their needs, rather than an undifferentiated mass of information. For discovering interesting and frequent navigation patterns from Web server logs many Web usage mining techniques have been applied. They propose semantically enriched Web Usage Mining method (SWUM), which combines the fields of Web Usage Mining and Semantic Web. In the proposed method, the undirected graph derived from usage data is enriched with rich semantic information extracted from the Web pages and the Web site structure. The results shows that proposed method is able to achieve 10 - 20% better accuracy than the solely usage based model, and 5-8% better than an ontology based model

Conclusion

In this paper we have discussed about important concept of Data mining i.e. Web Mining and also got some knowledge about Web Content Mining, Web Usage Mining, and Web Structure Mining. These are all sub concepts of Web Mining. In that Personalization in Web Usage Mining is upcoming research area. As World Wide Web usage increases E-Commerce applications and users of those applications are also increased. So there should be a system to get the current trends and fashions from the users. That should be updated periodically. For that purpose this Personalization or customization concept is used.

References

1. W. Bin and L. Zhijing, "Web Mining Research," in *Proceedings of the fifth International Conference on Intelligence and Multimedia Applications (ICCIMA'03),* 2003.
2. M. Eirinaki and M. Vazirgiannis, "Web Mining for Web personalization," *ACM Transactions on Internet Technology,* 2005.
3. A. J. Ratnakumar, "An Implementation of Web Personalization Using Web Mining Techniques," J*ournal of Theoretical and applied information technology*, 2005.

4. Girardi, R. and Marinho, L. B, "A domain model of web recommender systems based on usage mining and collaborative filtering", *Requirements Eng 12(1), 2007.*
5. D. Antoniou, M. Paschou, E. Sourla, and A. Tsakalidis, "A Semantic Web Personalizing Technique the case of bursts in web visits," presented at IEEE Fourth International Conference on Semantic Computing, 2010.
6. Pooja Mehtaa, Brinda Parekh, Kirit Modi, and Paresh Solanki, "Web Personalization Using Web Mining: Concept and Research Issue", *International Journal of Information and Education Technology, 2012*
7. Raymond Kosala and Henrick Blockel, "Web search: A Survey", SIGKDD explorations pages 95-104, july 2000.
8. Monika Soni, Rahul Sharma, Vishal Shrivastava, " Framework for Web Personalization using Web mining", IJRET, Volume 1, Issue 2, ISSN: 2319-1163.

IMAGE DENOISING USING VARIOUS WAVELET THRESHOLDING METHODS

K.Rajeswari, MCA., M.Phil.,
Assistant Professor, Department of Computer Science
Madurai Sivakasi Nadars Pioneer Meanakshi Women's College, Poovanthi

Abstract
The main aim of an image denoising *is to achieve both noise reduction and feature preservation. One goal in* image denoising *is to remove the noise from the* image *in such a way that the "original"* image *is discernible. A lot of combinations have been applied in order to find the best method that can be followed for denoising intensity images. There are several methods of noise removal from degraded images with Gaussian noise by using adaptive wavelet threshold (Bayes Shrink, Sure Shrink and VisuShrink) and compare the results in term of PSNR and MSE. Wavelet transforms enable us to represent signals with a high degree of scarcity. Wavelet thresholding is a signal estimation technique that exploits the capabilities of wavelet transform for signal denoising. This paper was to study various thresholding techniques such as SureShrink, VisuShrink and BayeShrink and determine the best one for image denoising.*
Keywords: *Denoising, wavelet thresholding, Bayes Shrink, SURE Shrink, VisuShrink.*

Introduction

Image denoising: Removing unwanted noise in order to restore the original image. Wavelet transform provides us with one of the methods for image denoising. Wavelet transform, due to its excellent localization property, has rapidly become an indispensable signal and image processing tool for a variety of applications, including denoising and compression. Wavelet denoising attempts to remove the noise present in the signal while preserving the signal characteristics, regardless of its frequency content. In many applications, image denoising is used to produce good estimates of the original image from noisy observations. The restored image should contain less noise than the observations while still keep sharp transitions (i.e. edges). Wavelet transform, due to its excellent localization property, has rapidly become an indispensable signal and image processing tool for a variety of applications, including compression and denoising. Wavelet denoising attempts to remove the noise present in the signal while preserving the signal characteristics, regardless of its frequency content. It involves three steps: a linear forward wavelet transform, nonlinear thresholding step and a linear inverse wavelet transform. Wavelet thresholding is a signal estimation technique that exploits the capabilities of wavelet transform for signal denoising. It removes noise by killing coefficients that are insignificant relative to some threshold, and turns out to be simple and effective, depends heavily on the choice of a thresholding parameter and the choice of this threshold determines, to a great extent the efficacy of denoising. Researchers have developed various techniques for choosing denoising parameters and so far there is no "best" universal threshold determination technique. Here we study various thresholding techniques such as SureShrink, VisuShrink and BayesShrink and determine the best one for image denoising.

Thresholding

Introduction

Threshold is a term which is not only applicable to image processing. In any field threshold has the same meaning. A threshold is a value which has two regions on its either side i.e. below the threshold or above the threshold. That is threshold means "cutoff value". In general any function can have a threshold

Hard and Soft Thresholding

Hard and soft thresholding with threshold, are defined as follows:

The hard thresholding operator is defined as:

D (U, λ) = U for all |U|> λ = 0 otherwise

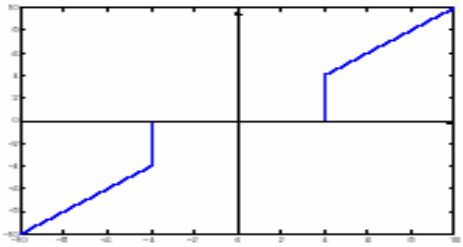


Figure 1: Hard Thresholding

The soft thresholding operator on the other hand is defined as:

D(U, λ) = sgn(U)max(0, |U| - λ)

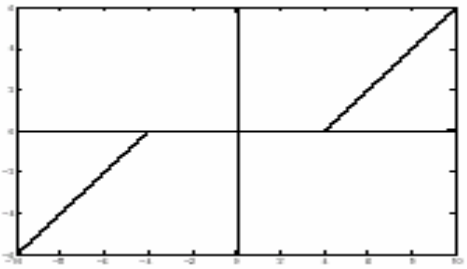


Figure 2: Soft Thresholding

Hard threshold is a "keep or kill" procedure and is more intuitively appealing. The transfer function of the same is shown in Fig 2.1. The soft thresholding (whose transfer function is shown in Fig.2.2), shrinks coefficients above the threshold in absolute value While at first sight hard thresholding may seem to be natural, the continuity of soft thresholding has some advantages. It makes algorithms mathematically more tractable. Moreover, hard thresholding does not even work with some algorithms such as the GCV procedure. Sometimes, pure noise coefficients may pass the hard threshold and appear as annoying 'blips' in the output. Soft thesholding shrinks these false structures.

Wavelet Thresholding

Wavelet thresholding (first proposed by Donoho) is a signal estimation technique that exploits the capabilities of wavelet transform for signal denoising. It removes noise by killing coefficients that are insignificant relative to some threshold.

- Types
- Universal or Global Thresholding
- Hard
- Soft
- SubBand Adaptive Thresholding

Image Denoising using Thresholding

The problem boils down to finding an optimal threshold such that the mean squared error between the signal and its estimate is minimized. The wavelet decomposition of an image is done as follows: In the first level of decomposition, the image is split into 4 subbands, namely the HH, HL, LH and LL subbands. The HH subband gives the diagonal details of the image; the HL subband gives the horizontal features while the LH subband represents the vertical structures. The LL subband is the low resolution residual consisting of low frequency components and it is this subband which is further split at higher levels of decomposition. The different methods for denoising we investigate differ only in the selection of the threshold.

The basic procedure remains the same:

- Calculate the DWT of the image.
- Threshold the wavelet coefficients. (Threshold may be universal or subband adaptive)
- Compute the IDWT to get the denoised estimate.

Soft thresholding is used for all the algorithms due to the following reasons: Soft thresholding has been shown to achieve near minimax rate over a large number of Besov spaces. Moreover, it is also found to yield visually more pleasing images. Hard thresholding is found to introduce artifacts in the recovered images. We now study three thresholding techniques- VisuShrink, SureShrink and BayesShrink and investigate their performance for denoising on the picture.

VisuShrink

VisuShrink is thresholding by applying the Universal threshold proposed by Donoho and Johnstone. This threshold is given by $\sigma\sqrt{2\log M}$ where σ is the noise variance and M is the number of pixels in the image. It is proved in that the maximum of any M values iid as $N(0,\sigma^2)$ will be smaller than the universal threshold with high probability, with the probability approaching 1 as M increases. Thus, with high probability, a pure noise signal is estimated as being identically zero. However, for denoising images, VisuShrink is found to yield an overly smoothed estimate as seen in Figure 4.11-4.14. This is because the universal threshold (UT) is derived under the constraint that with high probability, the estimate should be at least as smooth as the signal. So the UT tends to be high for large values of M, killing many signal coefficients along with the noise. Thus, the threshold does not adapt well to discontinuities in the signal.

Original Image	Noisy Image
	
Hard Thresholding Haar at level 1	Soft Thresholding Haar at level 1
	

Figure 1

Original Image	Noisy Image
	
Hard Thresholding db5 at level 1	Soft Thresholding db5 at level 1
	

Figure 2

Original Image	Noisy Image
	
Hard Thresholding db5 at level 2	Soft Thresholding db5 at level 2
	

Figure 3

Original Image	Noisy Image
	
Hard Thresholding db5 at level 3	Soft Thresholding db5 at level 3
	

Figure 4

SureShrink

SureShrink is a thresholding by applying subband adaptive threshold, a separate threshold is computed for each detail subband based upon SURE (Stein's unbiased estimator for risk), a method for estimating the loss $\|\hat{\mu} - \mu\|^2$ in an unbiased fashion. In our case let wavelet coefficients in the jth subband be { Xi : i =1,...,d }, $\hat{\mu}$ is the soft threshold estimator

$(\hat{X}_i) = \eta_t (X_i)$, we apply Stein's result to get an unbiased estimate of the risk

$E\| \hat{\mu}^{(t)} (x) - \mu \|^2 n$

$SURE(t;X) = d - 2\#\{i:|X_t| \le t\} \sum \min(|X_i|, t)^2$ i=1

For an observed vector x (in our problem, x is the set of noisy wavelet coefficients in a subband), we could find the threshold t^S that minimizes SURE (t: x),

$t^S = \text{argmin } SURE(t; X)$

The results obtained for the image on my picture using SureShrink are shown in Figure the 'Db5' wavelet coefficient was used with 3 levels decomposition. Clearly, the results are much better than VisuShrink. The sharp features of image are retained and the MSE is considerably lower. This because SureShrink is subband adaptive.

BayesShrink

BayesShrink is an adaptive data-driven threshold for image denoising via wavelet soft-thresholding. The threshold is driven in a Bayesian framework, and we assume Generalized Gaussian distribution (GGD) for the wavelet coefficients in each detail subband and try to find the threshold T which minimizes the Bayesian Risk. The results obtained by BayesShrink for my own picture is shown in Figure 5-8. The 'Db5' wavelet was used with three levels decomposition. We found that BayesShrink performs better than SureShrink in terms of MSE. The reconstruction using BayesShrink is smoother and more visually appealing than one obtained using SureShrink.

Adaptive Thresholding

Threshold Selection by Bayes Shrink and SURE Shrink Figure 5 & Figure 6

Original Image	Noisy Image
	
Bayeshrink with Haar at level 1	SureShrink with Haar at level 1
	

Figure 5

Original Image	Noisy Image
	
Bayeshrink with db5 at level 1	SureShrink with db5 at level 1
	

Figure 6

Original Image	Noisy Image
	
Bayeshrink with db5 at level 2	SureShrink with db5 at level 2
	

Figure 7

Original Image	Noisy Image
	
Bayeshrink with db5 at level 3	SureShrink with db5 at level 3
	

Figure 8

Experimental results

Following are the results of running the denoising algorithms for the methods discussed above on the picture. The denoising is done after adding the Gaussian noise with standard deviation 18 and mean 0 in the original picture. The MSE is calculated and compared for all the methods (Fig.5.1). The results are provided as a bar chart in the end as well they are tabulated.

The figures and MSE are compared for the global and adaptive thresholding techniques. The results are provided for thresholds selected by VisuShrink hard and soft methods and thresholds selected by default universal threshold for hard and soft thresholding. Similar thing is done with the BayesShrink and SureShrink methods.

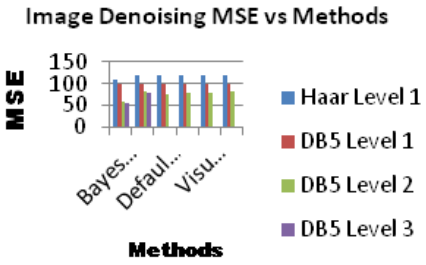


Figure 9: Comparison between all the methods based on MSE

Conclusion

We have seen that wavelet thresholding is an effective method of denoising noisy signals. We first tested hard and soft on noisy versions of the standard 1-D signals and found the best threshold. We then investigated many soft thresholding schemes such as VisuShrink, SureShrink and BayesShrink for denoising images. We found that subband adaptive thresholding performs better than a universal thresholding. Among these, BayesShrink gave the best results. It would be instructive to compare the performance of these algorithms on artificial images with discontinuities such as medical images.

References

1. Iain M.Johnstone David L Donoho. Adapting to smoothness via wavelet shrinkage.Journal of the Statistical Association, 90(432):1200–1224, Dec 1995.
2. David L Donoho. Ideal spatial adaptation by wavelet shrinkage. Biometrika,
3. David L Donoho. De-noising by soft thresholding. IEEE Transactions on Information Theory, 41(3):613–627,May 1995.
4. Maarten Jansen. Noise Reduction by Wavelet Thresholding, volume 161. Springer Verlag, USA,edition, 2001.
5. Martin Vetterli S Grace Chang, Bin Yu. Adaptive wavelet thresholding for image denoising and compression. IEEE Transactions on Image Processing, 9(9):1532–1546, Sep 2000.

A STUDY ON INTELLIGENT TRANSPORT SYSTEM

T.Ramaporkalai, M.C.A., M.Phil.,
Assistant Professor in CS Department
Madurai Sivakasi Nadars Pioneer Meenakshi Women's College, Poovanthi

Abstract
Transportation functions are necessary for any country's development. Road traffic congestion is a very big problem all over the world. After an extensive survey in the field of Vehicle and Highway System, different alternatives are analyzed to solve this problem and the concept of Intelligent Transportation System is proposed as the best solution. ITS bring significant improvement in transportation system performance, including reduced congestion and increased safety and traveler convenience. In this paper I have presented complete study of intelligent transport system.
Keywords: *Intelligent Transport system (ITS), Technologies, Applications*

Introduction

ITS [1] is acronym from Intelligent Transportation Systems. Under this name hides a collection of different technologies and management techniques which are used in public transport to improve efficiency of its In addition, these systems help to protect natural environment and improve the safety of road users. The name ITS was accepted on first world congress of this subject in 1994, Paris. The system was created in response to the growing environmental awareness, protesting against the negative effects of the automotive industry and also in response to the lack of expected effects of the constantly innovating investments that did not solve the problem of traffic systems bandwidth. As it turned, ITS is system without which is hard to imagine a smooth and safe movement 2/9 of vehicles in urban areas. Years of research conducted in the American and Canadian agglomerations shows that the use of these systems reduces costs of transport infrastructure 20% to 30%. In this paper I explained ITS architecture, Benefits of ITS, ITS Architecture, Challenges in ITS and various applications of Intelligent Transportation System.

ITs Architecture

The system architecture of the ITS [2] as shown in Figure:1 explains the data acquisition and evaluation technology, communication Network, digital mapping, video monitoring etc. This information helps in developing a system of traffic organization that enables information sharing the managers and users of traffic.

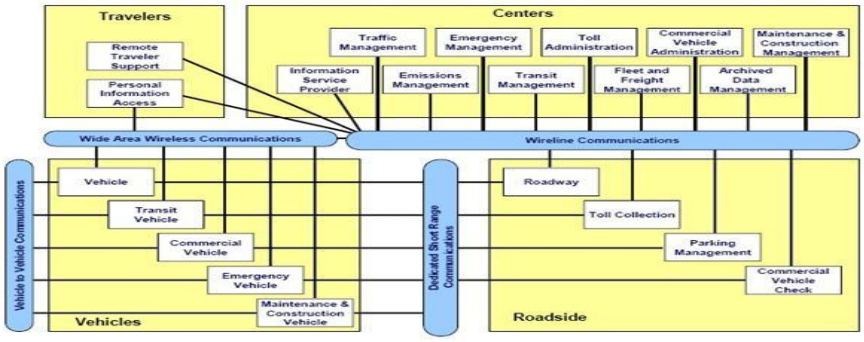


Figure 1 ITS Architecture

Benefits of ITS

ITS is viewed as an effort to channel technology applications to improve the performance of surface transportation systems in many ways. To this end it is expected to improve mobility (i.e., reduce congestion), reduce transportation-generated environmental impacts and energy consumption, enable improved safety, enhance quality of life including improved economic viability of communities, and increase the productivity of existing infrastructure.

ITS Applications

ITS are advanced applications aim to provide innovative services relating to transport. It is listed below.

Traffic Signal Control System(TSCS)

TSCS coordinate individual traffic signals to achieve wide traffic operations. These systems consist of intersection traffic signals, a communication Network, and a central computer or network of computers to manage the system. Coordination of traffic signals across the organizations requires the development of data sharing and signal control arrangements. Signal coordination systems are installed to provide access. Traffic control system is used to deliver the signal timing to drivers. In addition to control of traffic signals modern systems also provide wide range of surveillance capabilities, including various kinds of traffic detection and video surveillance.

In the advanced traffic control system [3] several adoptive control systems have been developed and deployed. These adoptive systems employ a methodology that use intersection detection data from dedicated sensor arrays to calculate optimum cycle length, splits and offsets which can react to unexpected traffic condition.

Traffic Signal Control with Connected Vehicle

A new initiative known as connected vehicles would allow for the wireless transmission of vehicles' positions, headings, and speeds to be used by the traffic controller. Instead of relying on point detectors (such as inductive loops or video detection systems) that sense only presence at fixed locations, signal systems would be able to use data from in-vehicle sensors transmitted wirelessly from equipped vehicles to the signal controller. Traffic signal control logic would have access to many measures that were previously estimated or unknown such as vehicle speeds, positions, arrival rates, rates of acceleration and deceleration, queue lengths, and stopped time.[4]

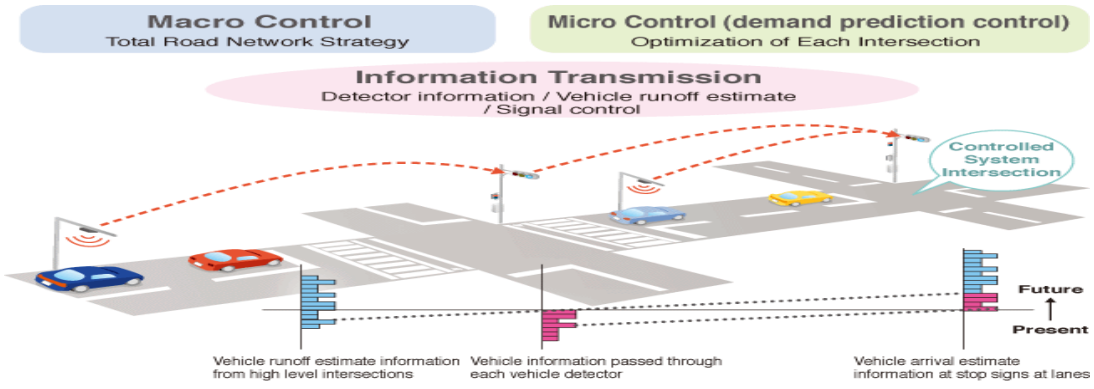


Figure 2 The outline for the traffic Signal Control System

Advanced Traveler Information Services (ATIS)

Advanced traveler information includes static and real-time information on traffic conditions, and schedules, road and weather conditions, special events, and tourist information [5]. It can be offered with value added options like sports scores, stock quotes, yellow pages and current news. ATIS is classified by how and when travelers receive their desired information (pretrip or en-route) and is divided by user service categories. Operations essential to the success of these systems are the collection of traffic and traveler information, the processing and fusing of information - often at a central point, and the distribution of information to travelers. Important components of these systems include new technologies applied to the use and presentation of information and the communications used to effectively disseminate this information.

Emergency Management Services

During Transport related emergencies, the use of ITS technologies can result in improved management of the emergency. ITS technologies provide transportation service and public safety agencies with the ability to communicate and coordinate operations and resources in real time. They support the data collection required for effective coordination of changing transportation system conditions and allow for the real-time implementation of operational and logistical strategies in cooperation with many partners. Efficient and reliable voice, data, and video communications further provide agencies with the ability to share information related to the status of the emergency, the operational conditions of the transportation facilities, and the location of emergency response resources.

Advanced Vehicle Control Systems(AVCS)

AVCS is viewed as an enhancement of the driver's control of the vehicle to make travel both safer and efficient. AVCS includes a broad range of concepts that will become operational on difference time scales. In the near term, collision warning systems would alert the driver to possible imminent collision. In more advanced system, the vehicle would automatically brake or steer away from the collision. Both systems are autonomous to the vehicle and provide substantial benefits by improving safety and reducing accident induced congestion. [6]

Wireless Traffic Signal Controller (Wi-TraC)

The WiTrac is a vehicle-actuated system that uses wireless technology to control traffic signals. The Master control of this system operates several sub-controls by sending wireless signals. The system is also equipped with cameras that constantly monitor traffic status and keep an eye on each and every vehicle within its jurisdiction. A special centralized control room monitors these cameras. The system is power efficient since it runs on solar power with power backup of up to 72 hours. Also its installation does not require digging up of roads to lay cables (it being wireless). The project has been developed at the cost of about Rs 14.75 Crore. CDAC claims that the system is highly power efficient, and the mounted solar panels offer the system 72 hours of backup time. [7]

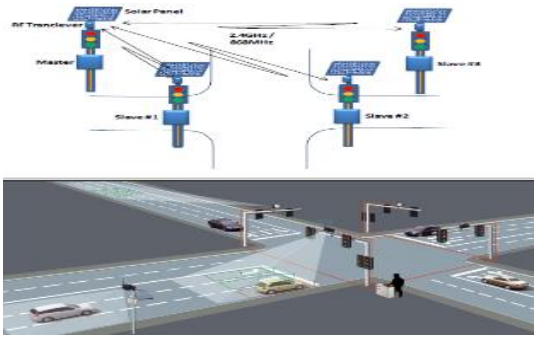


Figure 3:Wireless Traffic Signal Controller(Wi-Trac)

Advanced Public Transportation

Advanced public transportation systems (APTS) seek to apply transportation management and information technologies to public transit systems to increase their efficiency of operation and improve the safety of public transportation riders. Examples of APTS applications include real-time passenger information systems, automatic vehicle location systems, bus arrival notification systems, and systems providing priority of passage to buses at signalized intersections.

Technologies FOR ITS

ITS enabling many technologies to improve transportation conditions, safety and services. Some of them are listed.

Global Positioning System

The Global Positioning System (GPS) is the most common and accessible technique for vehicle localization. However, conventional localization techniques which mostly rely on GPS technology are not able to provide reliable positioning accuracy in all situations. This paper presents an integrated localization algorithm that exploits all possible data from different resources including GPS, radio-frequency identification, vehicle-to-vehicle and vehicle-to-infrastructure communications, and dead reckoning. A localization algorithm is also introduced which only utilizes those resources that are most useful when several resources are available. A close-to-real-world scenario has been developed to evaluate the performance of the proposed algorithms under different situations.[8]

Dedicated Short Range Communication(DSRC)

DSRC is dramatically increased for enhancing the road safety applications. The main task of DSRC is to protect the vehicles by communicating the warning message regarding the vehicle changing conditions, traffic occurrence and dangers over the road in the network. So, it is necessary to maintain the accurate communication timely with high reliability by implementing the appropriate protocol.[9]

Radio Frequency Identifier (RFID)

RFID technology has been known to be one of the noteworthy converging technologies of the 20th century. The technology can be applied in many fields. However, this paper focuses on the application of the technology in the transportation industry. The application of RFID in Intelligent Transport Systems (ITS) is gaining popularity with its widespread use in the field of toll management and the management of the overall transport sector. There are many RFID applications available in the market such as RFID contactless smart card commonly used in buses and LRTs, Automatic Vehicle Identification (AVI), Electronic Toll Collection (ETC), Smart Parking, and congestion zone pricing. In Mashhad, the second largest city of Iran, the "My Card" is used not only in Public transit but also in car parking, and soon in taxis and also other public municipality Services. Driven by such success stories, deployment of RFID technology in Mashhad is thus encouraged. This work has been carried out with a purpose to demonstrate benefits of the RFID technology in developing countries and its application in transport sector.[10]

Radio Model Communication (RMC)

Radio Modem Communication on UHF and VHF frequencies are widely used for short and long range communication within ITS. Short-range communications (less than 500 yards) can be accomplished using IEEE 802.11 protocols, specifically WAVE or the Dedicated Short Range

Communications standard being promoted by the Intelligent Transportation Society of America and the United States Department of Transportation. Theoretically, the range of these protocols can be extended using Mobile ad-hoc networks or Mesh networking. Longer range communications have been proposed using infrastructure networks such as WiMAX (IEEE 802.16), Global System for Mobile Communications (GSM), or 3G. Long-range communications using these methods are well established, but, unlike the short-range protocols, these methods require extensive and very expensive infrastructure deployment. There is lack of consensus as to what business model should support this infrastructure. [11]

Video Vehicle Detection (VVD)

The video vehicle detection (WD) algorithm based on virtual line is used extensively in intelligent transportation system (ITS). But it only utilizes luminance information of pixels, therefore its false reject rate (FRR) and false accept rate (FAR) are high. In order to solve this problem, an improved algorithm is proposed, which introduces two-level detection and utilizes both luminance and chrominance information. In the first level processing, it performs detection utilizing luminance information. In the second level processing, it does detection and modifies the luminance threshold utilizing chrominance information. Experiment results demonstrate that the improved algorithm can eliminate the influence of shadow and image noise effectively, and it is more accurate and robust.

Conclusion

This paper is a collection of basic concepts in Intelligent Transport System, architecture, benefits, applications and technologies used in ITS. Through ITS many areas take advantages such as data collection, toll collection, Traveler information and management, transit management etc. In future it is proposed to develop a prototype on "**Economy based Intelligent Transport Management system(EBITMS)"** and feasibility study on Indian Transportation.

References

1. Sen, Rijurekha, Vishal Sevani, Prashima Sharma, Zahir Koradia, and Bhaskaran Raman. "Challenges In Communication Assisted Road Transportation Systems for Developing Regions." NSDR, 2009 .
2. Campell, J.L., carney, C., and Kantowitz, B.H.., "Human factors Design Guidelines for Advanced traveler Information System(ATIS) and commercial Vehicle operation(CVO)", Federal Highway Admin., Mclean, VA, Rep. FHWA-RD-98-057-2, 2003.
3. Data Fusion for delivering Advanced traveler information services, U.S. Department of Transportation, Intelligent Transportation systems Joint program Office, Executive summary May 2003.
4. Traffic Signal with connected vehicle Published in Transportation Research Record: Journal of the Transportation Research Board, vol. 2381, 2013, pp. 65-72.
5. Intelligent Transportation Systems Field Operational Test Cross-Cutting Study, Advanced Traveler Information Systems September 1998 .
6. Perspective Intelligent system by Joseph S. Sussman 2008 - Technology & Engineering
7. http://indiannerve.com/cdac-witrac-indias-first-wireless-solar-powered-traffic-control-system-unveiled-9583/
8. Improving GPS-based vehicle positioning for Intelligent Transportation Systems- 2003
9. http:\www.ts.dot.gov/factsheets/dsrc_factsheet.htm
10. Applications and Opportunities for Radio Frequency Identification (RFID) Technology in Intelligent Transportation Systems: A Case Study IJIEE2013Vol.3 (3)341-345ISSN:2010-3719DOI: 10.7763/IJIEE.2013.V3.330.
11. [11][12] Improved video vehicle detection algorithm ieeexplore.ieee.org/ abstract/document/ 5536775.

DATA MINING: A REVIEW ON CLASSIFICATION ALGORITHMS

P.Priya
Assistant Professor, Department of Computer Science
Madurai Sivakasi Nadars Pioneer Meenakshi Women's College, Poovanthi

Abstract
Classification derives a model to determine the class of an object based on its attributes. A collection of records will be available, each record with a set of attributes. One of the attributes will be class attribute and the goal of classification task is assigning a class attribute to new set of records as accurately as possible. Classification is a supervised learning technique in data mining where training data is given to classifier that builds classification rules. Later if test data, is given to classifier, it will predict the values for unknown classes. Classification is process of generalizing the data according to different instances. Several major kinds of classification algorithms including SVM, C4.5, , Ada Boost, and Naïve Bayes and Apriori .This paper provides a inclusive survey of different classification algorithms and their advantages and disadvantages.
Keywords*- SVM, C4.5, Ada Boost, Naive Bayes, Apriori*

Introduction

Data mining refers to extracting or "**mining**" knowledge from large amounts of data. Knowledge Discovery in Data is the non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data. Data mining consists of more than collection and managing data; it also includes analysis and prediction. People are often do mistakes while analysing or, possibly, when trying to establish relationships between multiple features.

This makes it difficult for them to find solutions to certain problems. Machine learning can often be successfully applied to these problems, improving the efficiency of systems and the designs of machines. There are several applications for Machine Learning (ML), the most significant of which is data mining. Classification is a model finding process that is used for portioning the data into different classes according to some constrains. In other words we can say that classification is process of generalizing the data according to different instances. [1]

Classification Algorithm's In Data Mining

Support vector machines (svm)

The original SVM algorithm was invented by Vladimir N. Vapnik and the current standard incarnation (soft margin) was proposed by Vapnik and Corinna Cortes in 1995. In machine learning, **support vector machines** are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. More formally, a support vector machine constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks. Intuitively It is considered a good classifier because of its high generalization performance without the need to add a priori knowledge, even when the dimension of the input space is very high. The aim of SVM is to find the best classification function to distinguish between members of the two classes in the training data. The metric for the concept of the "best" Classification function can be realized geometrically. For a linearly separable dataset, a linear classification function corresponds to a separating hyper plane ***f(x)*** that passes through the middle of the two classes, separating the two. Once

this function is determined, new data instance x_n can be classified by simply testing the sign of the function $f(x_n)$; x_n belongs to the positive class if $(fx_n)>0$. Because there are many such linear hyper planes, SVM guarantee that the best such function is found by maximizing the margin between the two classes. Intuitively, the margin is defined as the amount of space, or separation between the two classes as defined by the hyperplane. Geometrically, the margin corresponds to the shortest distance between the closest data points to a point on the hyperplane. Having this geometric definition allows us to explore how to maximize the margin, so that even though there are an infinite number of hyperplanes,only a few qualify as the solution to SVM. To ensure that the maximum margin hyperplanes are actually found, an SVM classifier attempts to maximize the following function with respect to w and ***b:*** $L_p=1/2\ w \sum_{i=1}^{t} \alpha_i y_i(w .x_i +b)+ \sum_{i=1}^{t} \alpha_I$ where t is the number of training examples, and α_i **, i = 1, . . . ,** t**,** are non-negative numbers such that the derivatives of L_P with respect to αi are zero. α_i are the Lagrange multipliers and L_P is called the Lagrangian. In this equation, the vectors and constant b define the hyper plane. A learning machine, such as the SVM, can be modeled as a function class based on some parameters α . The VC dimension measures the maximum number of training examples where the function class can still be used to learn erectly, by obtaining zero error rates on the training data, for any assignment of class labels on these points. It can be proven that the actual error on the future data is bounded by a sum of two terms. The first term is the training error, and the second term if proportional to the square root of the VC dimension ***h.*** Thus, if we can minimize ***h,*** we can minimize the future error, as long as we also minimize the training error, SVM can be easily extended to perform numerical calculations. To extend SVM to perform regression analysis, where the goal is to produce a linear function that can approximate that target function. in support vector regression, or SVR, the error is defined to be zero when the differences between actual regressions are reported to be its insensitivity to outliers[6],[7].

C4.5 algorithm

C4.5 is an algorithm used to generate a decision tree developed by Ross Quinlan. C4.5 is an extension of Quinlan's earlier ID3 algorithm. The decision trees generated by C4.5 can be used for classification, and for this reason, C4.5 is often referred to as a statistical classifier. Nowadays C4.5 is renamed as J48 classifier in WEKA tool, an open source data mining tool. C4.5 builds decision trees from a set of training data in the same way as ID3, using the concept of information entropy. Training data is a set **S=s1,s2,....** of already classified samples. Each sample $\mathbf{S_i}$ consists of a p-dimensional vector $\mathbf{(x_1,i,x_2,i....,x_p,i)}$, where the $\mathbf{x_j}$ represent attributes or features of the sample, as well as the class in which $\mathbf{s_i}$ falls. At each node of the tree, C4.5 chooses the attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. The splitting criterion is the normalized information gain (difference in entropy). The attribute with the highest normalized information gain is chosen to make the decision. The C4.5 algorithm then recourses on the smaller sublists.

This algorithm has a few base cases.

- All the samples in the list belong to the same class. When this happens, it simply creates a leaf node for the decision tree saying to choose that class.
- None of the features provide any information gain. In this case, C4.5 creates a decision node higher up the tree using the expected value of the class.
- Instance of previously-unseen class encountered. Again, C4.5 creates a decision node higher up the tree using the expected value. [3]

Naive Bayes Algorithm

The Naive Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions. A more descriptive term for the underlying probability model would be "independent feature model". A naive Bayes classifier assumes that the presence or absence of a particular feature is unrelated to the presence or absence of any other feature, given the class variable. For some types of probability models, naive Bayes classifiers can be trained very efficiently in a supervised learning setting. In many practical applications, parameter estimation for naive Bayes models uses the method of maximum likelihood; in other words, one can work with the naive Bayes model without accepting Bayesian probability or using any Bayesian methods. Despite their naive design and apparently oversimplified assumptions, naive Bayes classifiers have worked quite well in many complex real-world situations. Bayes method—also called idiot's Bayes, simple Bayes, and independence Bayes. This method is important for several reasons. It is very easy to construct, not needing any complicated iterative parameter estimation schemes. This means it may be readily applied to huge data sets. It is easy to interpret, so users unskilled in classifier technology can understand why it is making the classification it makes. And finally, it often does surprisingly well: it may not Probabilistic approaches to classification typically involve modelling the conditional probability distribution $P(C|D)$, where *C* ranges over classes and *D* over descriptions, in some language, of objects to be classified. Given a description *d* of a particular object, we assign the class $argmax_c P(C = c|D = d)$. A Bayesian approach splits this posterior distribution into a prior distribution $P(C)$ and a likelihood

$$P(D|C):P(D = d|C = c)P(C = c)\ argmax_c P(C = c|D = d) = argmax_c\ p\left(D = \frac{d}{c} = c\right) p(C = c) \longrightarrow (1)$$

The denominator $P(D = d)$ is a normalising factor that can be ignored when determining the maximum *a posteriori* class, as it does not depend on the class. The key term in Equation (1) is $P(D = d|C = c)$, the likelihood of the given description given the class (often abbreviated to $P(d|c)$). A Bayesian classifier estimates these likelihoods from training data, but this typically requires some additional simplifying assumptions.

For instance, in an attribute-value representation (also called *propositional* or single-table representation), the individual is described by a vector of Values $a1 \ldots an$ for a fixed set of attributes $A1, \ldots, An$. Determining $P(D = d|C = c)$ here requires an estimate ofthe joint probability $P(A1 = a1, \ldots, An = an|C = c)$, abbreviated to $P(a1, \ldots, an|c)$. This joint probability
Distribution is problematic for two reasons:

(1) Its size is exponential in the number of attributes *n*, and
(2) It requires a complete training set, with several examples for each possible description. These problems vanish if we can assume that all attributes are independent

Given the class:

$$P(A1 = a1, \ldots, An = an|C = c) = \Pi^{n}_{i=1} P(Ai = ai|C = c) \longrightarrow (2)$$

This assumption is usually called the naive Bayes assumption, and a Bayesian classifier using this assumption is called the *naive Bayesian classifier*, often abbreviated to 'naive Bayes'. Effectively, it means that we are ignoring interactions between attributes within individuals of the same class. [2],[4].

Adaboost

Boosting is an approach to machine learning based on the idea of creating a highly accurate prediction rule by combining many relatively weak and inaccurate rules. The AdaBoost algorithm of Freund and Schapire was the first practical boosting algorithm, and remains one of the most widely used and studied, with application numerous fields.

Input: Data set $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \cdots, (x_m, y_m)\}$;
Base learning algorithm $\mathcal{L}$;
Number of learning rounds T.

Process:
$D_1(i) = 1/m$. % Initialize the weight distribution
for $t = 1, \cdots, T$:
$h_t = \mathcal{L}(\mathcal{D}, D_t)$; % Train a weak learner h_t from $\mathcal{D}$ using distribution D_t
$\epsilon_t = \Pr_{i \sim D_t}[h_t(x_i) \neq y_i)]$; % Measure the error of h_t
$\alpha_t = \frac{1}{2}\ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$; % Determine the weight of h_t
$D_{t+1}(i) = \frac{D_t(i)}{Z_t} \times \begin{cases} \exp(-\alpha_t) & \text{if } h_t(x_i) = y_i \\ \exp(\alpha_t) & \text{if } h_t(x_i) \neq y_i \end{cases}$
$= \frac{D_t(i)\exp(-\alpha_t y_i h_t(x_i))}{Z_t}$ % Update the distribution, where Z_t is
% a normalization factor which enables D_{t+1} be a distribution
end.
Output: $H(x) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$

Let X denotes the instance space and Y the set of class labels. Assume Y = {-1, +1}. Given a weak or base learning algorithm and a training set {(**x**1, y1), (**x**2, y2), . . . , (**x**m, ym)} where **x**i X and yi Y (i = 1, . . . ,m), the AdaBoost algorithm works as follows[15] First, it assigns equal weights to all the training examples (**x**i , yi) (i {1,2…….,m}). Let Dt Denote the distribution of the weights at the *t*-th learning round. From the training set and Dt the algorithm generates a weak or base learner ht :X ← Y by calling the base learning algorithm. Then, it uses the training examples to test ht, and the weights of the incorrectly classified examples will be increased. Thus, an updated weight distribution Dt+1 is obtained. From the training set and Dt+1 AdaBoost generates another weak learner by calling the base learning algorithm again. Such a process is repeated for T rounds, and the final model is derived by weighted majority voting of the T weak learners, where the weights of the learners are determined during the training process. In practice, the base learning algorithm may be a learning algorithm which can use weighted training examples directly; otherwise the weights can be exploited by sampling the training examples according to the weight distribution Dt.[9]

The Apriori Algorithm

Apriori algorithm is a one of the simple and famous data mining technique used for pulling out hidden patterns from data. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database. The frequent item sets determined by Apriori can be used to determine association rules which highlight general trends in the database: this has applications in domains such as market basket analysis. Apriori is designed to operate on database containing transactions. Other algorithms are designed for finding association rules in data having no transactions or having no time stamps transaction is seen as a set of items. Given a threshold C, the Apriori algorithm identifies the item sets which are subsets of at least C transactions in the database. Apriori uses a "bottom up" approach, where frequent subsets are extended one item at a time, and groups of candidates are tested against the data. The algorithm terminates when no further successful extensions are found. Apriori uses breadth-first search and a Hash tree structure to count candidate item sets efficiently. It generates candidate item sets of length K from item sets of length K-1. Then it prunes the candidates which have an infrequent sub pattern. According to the downward closurelemma,the candidate set contains all frequent K-length item sets. After that, it scans the transaction database to determine frequent item sets among the candidates. set of frequent itemsets of size *k* be *Fk* and their candidates be *Ck.*

F_l=(Frequent itemsets of cardinality 1);
for($k = 1; F_k \neq \phi; k++$) do begin
C_{k+1} = apriori-gen(F_k); //New candidates
for all transactions $t \in$ Database do begin
C'_t = subset(C_{k+1}, t); //Candidates contained in t
for all candidate $c \in C'_t$ do
$c.count++$;
end
$F_{k+1} = \{C \in C_{k+1} \quad | c.count \geq$ minimum support $\}$
end
end
Answer $\cup_k F_k$;

The pseudo code for the algorithm is given below for a transaction database T, and a support threshold of Є. Usual set theoretic notation is employed though note that T is a multiset. C_k is the candidate set for level K. Generate() algorithm is assumed to generate the candidate sets from the large item sets of the preceding level, heeding the downward closure lemma. Count[c] accesses a field of the data structure that represents candidate set C, which is initially assumed to be zero[5].

Advantages and Disadvantages of Classification Algorithms

S.No	Algorithms	Advantages	Disadvantages
1	Support vector machine Algorithm	1. Produce very accurate classifiers. 2. Less overfitting, robust to noise 3.Especially popular in text classification problems where very high-dimensional spaces are the norm. 4. Memory-intensive .	1.SVM is a binary classifier. To do a multi-class classification, pair-wise classifications can be used (one class against all others, for all classes). 2. Computationally expensive, thus runs slow.
2	C4.5 Algorithm	1.It produce the accurate result 2.It takes the less memory to large program execution. 3.It takes less model build time. 4.It has short searching time	1.Empty branches. 2.Insignificant branches. 3.Over fitting.
3	Navie bayes Algorithm	1. Simplicity. 2.Computational efficiency. 3Good performance. 4.Spam filteri	1.The naive Bayes classifier requires a very large number of records to obtain good results. 2. A predictor category is not present in the training data, naive Bayes assumes that a new record with that category of the predictor has zero probability.
4	Adaboost Algorithm	1.Very simple to implement 2.Does feature selection resulting in relatively simple classifier 3.Fairly good generalization	1. Suboptimal solution. 2.Sensitive to noisy data and outliers
5	Apriori Algorithm	1.Simplicity. 2.Support (utility). 3.Confidence (certainty). 4.Algortihm uses knowledge from previous iteration phase to produce frequent itemsets .	1.Needs several iterations of the data 2.Uses a uniform minimum support threshold 3.Difficulties to find rarely occuring events 4.Alternative methods can address this by using a non-uniform minimum support thresold 5.Some competing alternative approaches focus on partition and sampling

Conclusion

This paper presents the detailed description of the classification algorithms, and their merits and demerits, these classification algorithms can be implemented on different types of data sets like data of patients, financial data according to performances. On the basis of the performance of these algorithms, these algorithms can also be used to detect the natural disasters like cloud bursting, earth quake, etc.

References

1. J. Han, and M. Kamber, 2006. Data Mining Concepts and Techniques, Elsevier Publishers.
2. Anshul Goyal, Rajni Mehta, "PerformanceComparison of Naive Bayes and J48 ClassificationAlgorithms", IJAER, Vol. 7, No. 11, 2012, pp.
3. Mohd. Mahmood Ali1, Mohd. S. Qaseem, Lakshmi Rajamani,and A. Govardhan," EXTRACTING USEFUL RULES THROUG IMPROVED DECISION TREE INDUCTION USING INFORMATION ENTROPY", (IJIST) Vol.3, No.1, January 2013.
4. Vidhya.K.A, G.Aghila" *A Survey of Naïve Bayes Machine Learning approach in Text Document Classification*" *(IJCSIS) Vol. 7, No.2,2010*
5. Charanjeet Kaur, "Association Rule Mining using Apriori Algorithm: A Survey", IJARCET Volume 2, Issue 6, June 2013.Anshul Goyal, Rajni Mehta, "Performance Comparison of Naive Bayes and J48 ClassificationAlgorithms", IJAER, Vol. 7, No. 11, 2012, pp.
6. Shigeo Abe., Support Vector Machines for Pattern Classifications, Second Edition. Springer, Londres; 2010.
7. J. Smola, B. Scholkopf, " A tutorial on support vector regression", Stat Comput 14 (2004) 199–222.
8. H. Bhavsar, A. Ganatra, "Variations of Support Vector Machine Classification : A survey", International Journal of Advanced Computer Research, Volume 2, Number 4, Issue 6 (2012) 230–236.
9. Raj Kumar,Dr. Rajesh Verma," Classification Algorithms for Data Mining P: A Survey" IJIET Vol. 1 Issue August 2012, ISSN: 2319 – 1058.
10. Xindong Wu et.al, "Top 10 Algorithms of Data Mining", Springer-Verlag London**,** 2007.

VIGOR EFFECTIVE VIDEO COMPRESSION MECHANISM FOR WIRELESS SENSOR NETWORK USING EDGE FEATURE REDUCTION ALGORITHM

S.Samera Banu

Assistant Professor, Madurai Sivakasi Nadars Pioneer Meenakshi Women's College, Poovanthi

Abstract

With the growth of multimedia and internet, compression techniques have become the thrust area in the fields of computers. Popularity of multimedia has led to the integration of various types of computer data. Multimedia combines many data types like text, graphics, still images, animation, audio and video. Video compression is a process of efficiently coding digital image to reduce the number of bits required in representing image. Its purpose is to reduce the storage space and transmission cost while maintaining good quality. Many different image compression techniques currently exist for the compression of different types of videos. In the present research puts forward a low-complexity video compression algorithm that uses the edges of objects in the frames to estimate and compensate for motion. We conduct extensive computational simulations to examine the truth of our method and find that the proposed schemes not only balance energy consumption of sensor nodes by sharing of the processing tasks but also improve the quality of decoding video by using edges of objects in the frames.

Keywords: *Edge Detection, Video Compression, WSN, Energy constraints, Color Model Separation.*

Introduction

A Wireless Video Sensor Network (WVSN) is a special kind of Wireless Sensor Network (WSN) that is capable of capturing video data at video sensor nodes. Two types of nodes are considered in WVSNs, Video Sensor Nodes (source nodes) and Processing Sensor Nodes that affect the retrieval of video data. Use of high rate Wireless Video Networks could include on-land and at sea surveillance, video assisted navigation, video assisted ship management and remote monitoring of training exercises. Since the Wireless Sensor Devices in general have very limited resources in terms of energy, computation, communication, and memory they demand a lot of the video compression algorithms. This video compression algorithm must have low-complexity to decrease energy drain at battery constrained sensors. The decode estimates motion (motion vectors) and compensates for motion based on active regions and reference frame. Decompression has done by using inverse transform.

In this paper we proposed a video compression algorithm in which video are converted into frames, and the edges are detected by Homogeneity Edge Detection. This edge is concatenated to find Mark Active Region which is used for video compression by Adaptive Arithmetic Coding (AAC).Decompression is achieved by reverse Discrete Wavelet Transform. Then the Analytic parameter such as PSNR, Compression Rate is computed and it is plotted.

Homogeneity Edge Detection

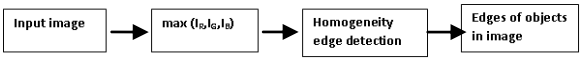


Figure 1.Homogeneity edge detection method

In this we modify the conventional homogeneity edge detection method by inserting the preprocessing block before detecting edges of objects in images to reduce the complexity computation. Therefore, the detection process includes two steps as follows.

Step 1 (preprocessing). In this step, we select one of three color elements of input image to detect as follows.

$I(i,j)=\max(I_R(i,j),I_G(i,j),I_B(i,j))$….(1)

I_R (i,j)-Intensity of red element of input image.

I_G (i,j)-Intensity of green element of input image.

I_B (i,j)-Intensity of blue element of input image.

The goal of the step is to reduce energy consumption of wireless sensor devices for detecting edges of objects by carrying out edge-detection task one time.

Step 2 (detecting edges of objects). We use the homogeneity operator that performs to calculate the different value of center point with eight neighbors to find edges of objects in image (E_A), as shown in Figure 1. The homogeneity operator is defined as

$$E_A(i,j)=\begin{cases}\max\{|I_A(I,j)-I_{An}(I,j)|n=1,2,\ldots\\ \text{if } \max\{|I_A(i,j)-I_{An}(i,j)|\}\geq threshold_1\\ 0, \text{otherwise} \ldots\ldots\ldots\ldots\ldots (2)\\ 0\leq threshold_1\leq 255\end{cases}$$

Where threshold1 is the threshold that is used to improve the quality of edges of objects in image

Proposed Video Compression Algorithm

We propose a method to find motion regions based on comparing the edges of objects among frames. From references, the proposed algorithm has three different points. First, we use the difference of edges of objects among frames to mark motion regions while uses the differences among noise, shadow and illumination regions [1]. Secondly, we use edges of objects in the background images to increase accuracy when performing to mark motion regions [2]. Thirdly, we then apply our method (finding the motion regions by comparing the edges of objects) for MPEG-2/H.262 encoder that is suitable to perform for wireless applications because of its low complexity of algorithm and acceptable quality of decoding data [3]. The main difference between the proposed algorithm and MPEG-2 is that only the edges of objects in the frames are required in the proposed method while MPEG-2 requires all data of the frames. Therefore, we can save energy and time for finding motion vectors and compensating for motion.

Figure 2 depicts an overview of the proposed video compression system. First, the source node captures the current frame to detect its edges of objects and compares the detected edges of objects with the edges of objects in the background images. The goal of this step is to reduce noise. Then the detected edges of objects of this frame are stored in the buffer of source node. The source node repeats the same process for the next frame and compares the detected edges of objects of this frame with those of the previous frame in its buffer to mark active regions. Based on the active regions, the source node finds motion vectors and compensates for motion. Finally, the motion regions are transformed, quantized, run length encoding (RLE), and Adaptive Arithmetic coding respectively, and the motion vectors are encoded by RLE, and Adaptive arithmetic coding by the encode block.

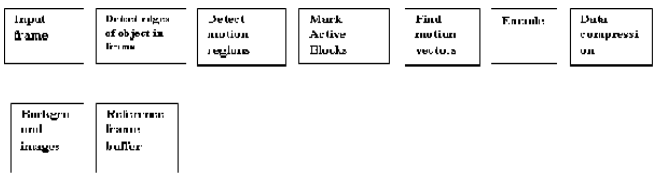


Figure 2 The proposed video compression algorithm

Mark Active Regions

First, the edge-detector block performs to detect edges of objects in the frames ($P1$, $P2$ frames, and background images). At the initial time, we assume that the background images were stored in the buffer. At the next times, we will use the previous frames as the background images. The detected edges of objects among frames are then compared each other frames. Secondly, (marking the active regions), the number of pixels ($N\Delta P12$) whose value is different from zero in each block of $\Delta P12$ frame is calculated. If $N\Delta P12$ is more than threshold2, which depends on the size of block, the block is marked. In our simulation, threshold2 is set up as 32 if the size of block is 8 × 8.

The algorithm then performs to find motion vectors at the motion estimation block and compensate for motion at the motion compensation is based on the marked regions (active regions). To save energy consumption for finding motion vectors, we use a "three-step search" algorithm. The algorithm searches motion vector based on comparing and determining minimum mean absolute errors (MAEs) of eight points that have the same distance from center point (estimated point). From figure 3, frame 1, frame 2 and back ground image are transferred to the edge detector block and to mark active regions [4].

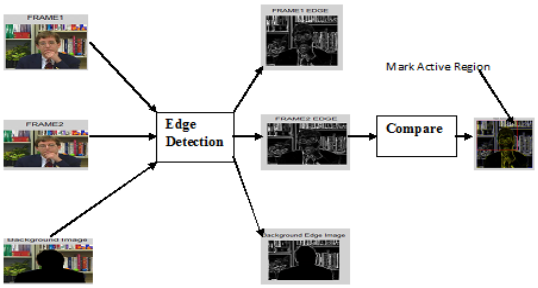


Figure 3 Comparing the edges of objects in frames to mark active regions

Finally, the motion regions are transformed DCT, quantized, RLE, and Adaptive Arithmetic encoding, respectively, and motion vectors are encoded by RLE and Adaptive arithmetic coding. Based on the proposed video compression algorithm, we propose two different schemes to implement this video compression algorithm in Wireless Video Sensor Networks (WVSNs).

Encoding

Encoding System takes an original image as input and after processing on this, it gives compressed image as output. While Decoding System takes a compressed image as input and gives the image as output which is more identical to original image. Wavelet analysis can be used divided the information of an image into approximation and detailed Sub signal. The approximation sub signal shows the general trend of pixel value, and three detailed sub signal show vertical, horizontal and diagonal details or changes in image. If these detail is very small than they can be set to zero without significantly changing the image. If the number of zeros is greater than the compression ratio is also greater. There is two types of wavelet is used. First one is Continues wavelet transform and second one is Discrete wavelet transform .Wavelet analysis is computed by filter bank. There is two type of filter

- **High Pass Filter:** high frequency information is kept, low frequency information is lost.
- **Low Pass Filter:** low frequency information is kept, high frequency information is lost. So signal is effectively decomposed into two parts, a detailed part (high frequency) and approximation part (low frequency). Level 1 detail is horizontal detail, level2 detail is vertical detail and level3 detail is diagonal detail of the image signal. In algorithm there is shown one level discrete wavelet transform. You can also increase the level of DWT by applying this process more than one time. Second and third level DWT gives the better compression ratio. But it will come with loss of some information. First level DWT is quite reasonable for both achieving high compression ratio and also got quality. In numerical analysis and functional analysis, a discrete wavelet transform (DWT) is

any wavelet transform for which the wavelets are discretely sampled. As with other wavelet transforms, a key advantage it has over Fourier transforms is temporal resolution: it captures both frequency and location information (location in time).

Adaptive Arithmetic Coding

In order to encode images:

- Divide image up into 8x8 blocks.
- Each block is a symbol to be coded.
- Compute Adaptive Arithmetic codes for set of block
- Encode blocks accordingly.

Unlike other types of compression, in Arithmetic coding a sequence of n symbols is represented by a number between 0 and1.Arithmetic coding came from Shannon's that sequences of symbols can be coded by their cumulative probability. From the decoding algorithm, we can see that as the interval is divided, the number of binary sequence also doubles. Therefore we can say that for an arbitrary code range, lN, the minimum encoding length, Lmin=@log2 (lN) …….. (3)

For an encoding sequence S, the number of bits per symbol is bounded by: L S≤ symbol bits NlN/) (log2−σX where σis the total compression overhead including bits required for saving the file, bits representing number of symbols, and information about the probabilities. The average number of bits approaches entropy. W e can see that arithmetic coding achieves optimal performance. As with Adaptive Arithmetic coding (AAC) also reduces the number of passes through the data from two to one. The difference between the two is that there is no need to keep a tree for the code words. The only information that needs to be synchronized is the frequency of occurrence of the symbols.

Edge Feature Reduction

The inability of state of the art approaches in color image processing to properly segment objects from the background without interference from object shadows and highlights. The inability of existing color similarity measures to evaluate directly color differences based on hue without considering intensity in RGB Here we are providing a new method in video compression schemes using edge feature on wireless video sensor networks. The proposed algorithm has three different points from. First, we use the difference of edges of objects among frames to mark motion by color edge detection algorithm. Since we use edges of objects, processing time and energy consumption for encoding frames of our method in Adaptive Arithmetic Coding Secondly, we use edges of objects in the background images to increase accuracy when performing to mark motion regions. As a result, the numbers of motion vectors and motion regions reduce, and thus compression rate is improved. Thirdly, we then apply our method to finding the motion regions by comparing the edges of objects.

Simulation Results

We have used Math lab R2009B for our simulation, and we evaluate three parameters, the quality of decoding video (compression rate, video quality, and encoding time), the quality of the network (the numbers of received and lost frames), and the power consumption of the network. To evaluate the video compression algorithms, three parameters, PSNR, compression rate, and encoding time of decoding video, are used. Image quality is measured by PSNR metric as the following equation:

PSNR [db] = 20log10 2^b-1

E||x_1 (i,j)-x_2 (i,j)|| (4)where $x_1(i, j)$ is the value of pixel (i, j) in the original image, $x2(i, j)$ is the value of pixel in the decompressed image, and b is the number of bits per pixel for the original image.

The storage capacity is measured by the compression rate. It is defined as follows:

Compression rate = Size of encoded image / Size of original image

The PSNR and the compression rate have relation with each other. If the compression rate metric decreases, the image quality will go down. It means that PSNR will decrease. Thus, we need to balance two parameters. When compressing image, while researchers focus on the compression rate metric because of the energy constraints. Here, the PSNR value, Compression rate are considered.

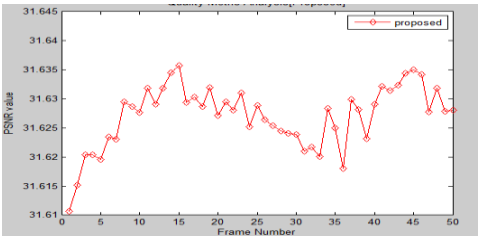


Figure 4 Simulation Result using Math lab

Figure shows the plotting of PSNR value versus frame number of the video. It shows the PSNR values for all the 50 frames of the video.

Conclusion

Our proposed video compression schemes using an edge detection technique for balancing energy consumption in WSNs. The proposed schemes solved three problems on WSNs, energy conservation, resource constrained computation, and data processing. The energy conservation and data processing problems are solved by distributing the data processing tasks for multiple nodes from a source node to the cluster head in a cluster. As a result, the number of received data by the proposed schemes increases while the energy consumption is just the same for all schemes and energy among sensor nodes is balanced. For the resource-constrained computation problem, we use edge feature of image to find motion regions. The advantages of the technique are short execution time, low computational complexity, and low error rate.

References

1. Seema and M. Reisslein, "Towards efficient wireless video sensor networks: a survey of existing node architectures and proposal for a flexi-wvsnp design," *IEEE Communications Surveys and Tutorials*, vol. 13, no. 3, pp. 462–486, 2011.
2. H. Abid and S. Qaisar, "Distributed video coding for wireless visual sensor networks using low power Huffman coding," in *Proceedings of the 44th Annual Conference on Information Sciences and Systems (CISS '10)*, pp. 1–6, March 2010.
3. N. Ling, "Expectations and challenges for next generation video compression," in *Proceedings of the 5th IEEE Conference on Industrial Electronics and Applications (ICIEA '10)*, pp. 2339–2344, June 2010.
4. U. Datta and S. Kundu, "Packet level performance of a CDMA wireless sensor networks in presence of correlated interferers," in *Proceedings of the 4th International Conference on Computers and Devices for Communication (CODEC '09)*, pp. 1–4, December 2009.

COMPARISON OF SQL AND NOSQL DATABASE THROUGH BIGDATA

D.Nivetha, M.Sc., M.Phil.,
Madurai Sivakasi Nadars Pioneer Meenakshi Women's College, poovanthi

Abstract
In a recent advance technique cloud computing and distributed web applications has created the need to store large amount of data in distributed databases that has high availability and scalability. In a modern world growing number of companies have adopted various types of non-relational databases, commonly referred to as NoSQL databases, and their application has extensive market interest. In the current emphasis on "Big Data", NoSQL databases have surged in popularity. These databases are claimed to perform better than the SQL databases. In this paper, it can analyze the SQL and NOSQL database in the base of data stores comparison. MongoDB is one of the NOSQL databases.
Keywords: *NOSQL, MongoDB, Big data, SQL*

Introduction

Structured Query Language (SQL) databases are table-oriented. Each table has a predefined structure as part of the schema. NOSQL is an unstructured Query Language and data stores are widely used to store and retrieve possibly large amounts of data, typically in a key-value format. Databases including HBase, MongoDB and Cassandra are more popular, for their capability in dealing with large amount of complex data in various structures. This led to the development of a new data model in NOSQL.

NoSQL is originally referring to "non SQL", "non relational" or "not only SQL" database, which provides mechanism for storage and retrieval of data in means other than the tabular relations used in relational databases. NoSQL databases were coined in 1998 and it tries to make data model without using SQL. NoSQL database provide facility for reading, writing database contents and storage becomes much easier and quicker than the traditional SQL.

SELECT, UPDATE, DELETE are the commonly used data manipulation statements. Though the concept of NOSQL was developed a long time ago, it was after the introduction of database as a service (DBaaS), it gained a prominent recognition. Because of the high scalability provided by NOSQL, it was seen as a major competitor to the traditional relational database model. In MongoDB, it is only a small fraction of time taken for query execution.

MongoDB is an open-source database and stores data in Java Script Object Notation (JSON). JSON documents vary in structure. Related information can be stored together for fast query access through the MongoDB query language. MongoDB is designed with high availability, scalability, replication and auto-shrading. MongoDB is more suitable for applications having large volume of data with high security. Since it is schema-less, it supports different types of data. Most NOSQL systems have removed the multi-platform support and some extra unnecessary features of RDBMS, making them much more lightweight and efficient than their RDMS counterparts. The NOSQL data model does not guarantee ACID properties (Atomicity, Consistency, Isolation and Durability) but instead it guarantees BASE properties (Basically Available, Soft state, Eventual consistent).

Our paper is organized as follows, Section I deals with Introduction, Section II analyzes various literatures, while Section III focus on proposed system and Section IV presents the conclusion.

Litertature Survey

The literature survey analyzes NOSQL database with MongoDB which is a cross-platform document -oriented database system, characterized by mass data storage and good query performance [9]. The non expert database users cannot access the database using SQL*Plus [8]. The advantage of NOSQL databases work well with distributed data stores such as Google and Facebook, where large data types are handled and need to be stored [4].

Non-relational databases can be used as a complement to relational databases, increasing its performance and bringing different characteristics and advantages. The relational databases allow database to store and extract retrieval of data using standard SQL language. Until it has some limitation in data storage, efficiency and losing of query in large amount of data set values. To solve this problem, a new model has been developed with large number of features called as NOSQL [7]. In spite of reducing query to execute complex as well as simple queries has proved to be an efficient system to people, who do not have much knowledge about databases [3].

Currently, there are over 150 NOSQL databases with diverse features and optimizations [2]. A model in [1] helps to represent hierarchical relationships to store arrays and other complex structures very easily. Documents in a record need not have an identical set of fields. In MongoDB, the speed gives a good performance, scalability reduces work load by increases resource pool. MongoDB is the best solution for their evaluation purpose because it is able to handle less data along with large data proficiently [6]. The value string consists of double quotes like “field name” followed by colon “:” and then it can be separated comma with next field [5]. The values can be other documents, arrays and array of documents and each pair is separated by comma. Any relational database has a typical schema design that shows number of tables and the relationship between these tables. While in MongoDB, there is no concept of relationship. It is non-relational database with dynamic schema.

Proposed System

The traditional database system maintained are the relational database systems, in which all the database retrieval and transactions are made through Structured Query Language (SQL). Since the database becomes huge in size, we need to have transition from normal database retrieval to big data retrieval. This paradigm requires a NoSQL database, MongoDB.

The necessity of using MongoDB is to have document oriented storage. The provision for indexing any attribute and usage of rich queries and fast in-place updates makes MongoDB more attractive. MongoDB can be applied in the field of Big Data, Content Management and Delivery, Mobile and Social Infrastructure, User Data Management and Data Hub.

The following terminologies are vital in MongoDB,

Database in MongoDB, is a physical container for collections. Each database gets its own set of files on the file system. A single MongoDB server typically has multiple databases [10].

Collection is a group of MongoDB documents. It is the equivalent of an RDBMS table. A collection exists within a single database and do not enforce a schema. Documents within a collection can have different fields. Typically, all documents in a collection are of similar or related purpose. A document is a set of key-value pairs. Documents have dynamic schema, which mentions documents in the same collection and do not need to have the same set of fields or structure.

Distributed Databases

In distributed database system models, data is logically integrated data storage and processing is physically distributed across multiple nodes in a cluster environment. Distributed database is a collection of multiple nodes connected together over a computer network and act as a single point to the user. Advantages of distributed database system include hardware scalability, replication of data across nodes in the cluster, concurrent transactions, availability in case of node failure database will still be online and performance improvements because of distributed hardware.

NOSQL Databases

Atomicity: All of the operations in the transaction will complete, or none will.
Consistency: Transactions never observe or result in inconsistent data.
Isolation: The transaction will behave as if it is the only operation being performed
Durability: Upon completion of the transaction, the operation will not be reversed.

NOSQL Movement

The high volume amount of data in the web is a problem which has to be considered by successful web pages like the ones of Facebook, Amazon and Google. Besides dealing with tera and pet bytes of data, massive read and write requests have to be responded without any noticeable latency. In order to deal with these requirements, these companies maintain clusters with thousands of commodity hardware machines. Due to their normalized data model and their full ACID support, relational databases are not suitable in this domain, because joins and locks influence performance in distributed systems negatively.

In addition to high performance, high availability is fundamental requirement of many companies. Therefore, databases must be easily replicable and have to provide an integrated failover mechanism to deal with node or datacenter failures.

CAP Theroem

Consistency: The data is always the same in every replication on every server [11].
Availability: The availability of data can be accessible in a various ways.
Partition Tolerance: In this database it works fine despite network and machine failures. The theorem says that only two of these aspects can be guaranteed at the same time in a distributed system. You have to "pick" two of them. In this paper, we will not discuss the proof of this theorem; we will just accept it as a matter of fact [12].

Sharding

It is a partitioning mechanism in which records are stored on different servers according to some key. The data is partitioned in such a way that records, that are typically accesses/updated together, reside on the same node. The load is almost evenly distributed among the servers. Some systems also use vertical partitioning in which parts of a single record are stored on different servers.

Comparison of Structured and Unstructured Model

Structured data sets are those where the activity of processing and output is predetermined and highly organized. Structured systems are designed. Payroll, Inventory control systems, point of sale systems, airline reservations are all forms of structured systems since they are using structured data-the data which is stored and displayed as a set of rows and tables. In contrast, unstructured data sets

are the data that have little or no predetermined form or structure. Unstructured data sets include email, contracts, blogs, and other communications. A person who performs a communications activity in an unstructured system has wide latitude to structure the message in whatever form is desired. The rules of unstructured systems are fewer and less complex [15, 16]. The structured and unstructured data can be different from technical, organizational, structural, functional point of view

SQL	NOSQL
Table	Collection
Row, Column	Document field
Index	Index
Table Joins	Embedded Documents
Fixed Schema	Schema less
Scalability is low	High Scalability
Supports ACID Property	Support ACID & BASE property

Table1 Comparison of SQl and NOSQ

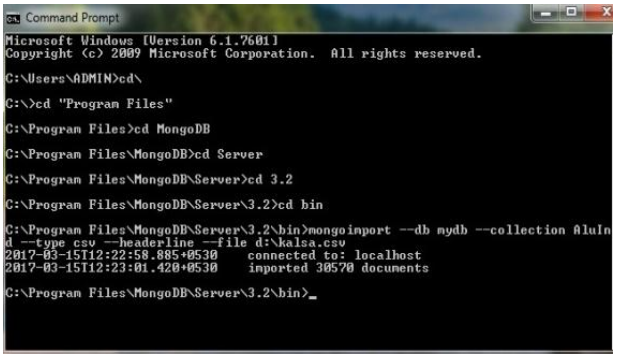


Figure 1 Imported Document in NoSQL databases

Conclusion

As we are in the Big data era, the expectation level in data storage and retrieval processing is much more expected than ever. The powers of NOSQL databases are tremendous in terms of processing and retrieval. In the past, data was once considered too expensive to store. But now in the present scenario, huge data can be captured, stored and processed through NoSQL databases. This performance is suitable for deciding which database will be used for enterprises and applications. The retrieval processing has ended up in faster transaction through their operation insertion, updation, deletion, sorting and import. Advantage of using NOSQL lies in internal memory, for storing the working set, which in turn enable faster access of data. It has quite good performance and it doesn't have several constraints and restrictions as compared to other NoSQL databases.

References

1. Dipina Damodaran, Shirin Salim and Surekha Marium Vargese, " Performance Evaluation of mysql and MongoDB databases", International Journal on Cybernetics & Informatics (IJCI) Vol. 5, No. 2, April 2016.
2. Harpinder Kaur, Janpreet Singh, " Improvement in Load Balancing Technique for MongoDB Clusters", International Journal of Applied Information Systems Vol 8– No.4, February 2015.
3. Mohammad Farhan, Yatin Chauhan, Mohammad Akhtar, Fayeem Khan, Poonam Pangarkar, "Object Query Optimization through Detecting Independent Sub queries", IOSR Journal of Computer Engineering (IOSR-JCE) Vol 16, Issue 2, April 2014.
4. Ms. Neetu Sharma, Ms. Charu Jain, Mr. Mahesh Chauhan, "Comparative Study of Distributed, Scalable, & High Performance NoSQL databases", IJITE Vol.03 Issue-01, January 2015.
5. Rupali Arora, Rinkle Rani Aggarwal, "Modeling and Querying Data in MongoDB", International Journal of Scientific & Engineering Research", Vol 4, Issue 7, July-2013.
6. Veronika Abramova1, Jorge Bernardino, and Pedro Furtado, "Experimental Evaluation of NoSQL Databases", International Journal of Database Management Systems (IJDMS) Vol.6, No.3, June 2014.
7. Veronika Abramova A, Jorge Bernardino A, B, Pedro Furtado B, "Which NoSQL Database? A Performance Overview", Open Journal of Databases (OJDB) Vol 1, Issue 2, 2014.

8. S.Vijayprasath et al, "International Journal of Computer Science and Mobile computing", IJCSMC, Vol. 4, January 2015.
9. Yogesh Punia, Rinkle Aggarwal, "Implementing Information System Using MongoDB and Redis", International Journal of Advanced Trends in Computer Science and Engineering", Vol. 3, March 10, 2014.
10. http://www.tutorialspoint/MongoDB.com
11. Mohamed A. Mohamed Obay G. Altrafi Mohammed O. Ismail , "Relational vs. NoSQL Databases: A Survey", International Journal of Computer and Information Technology (ISSN: 2279 – 0764) Volume 03 – Issue 03, May 2014
12. S. Weber, "NoSQL Databases," University of Applied Sciences HTW Chur, Switzerland, 2010.
13. N. A. L. Seth Gilbert,"Perspectives on the CAP Theorem," Singapore, 2012.
14. Sanobar Khan, Prof.Vanita Mane,"SQL Support over MongoDB using Metadata", International Journal of Scientific and Research Publications, Volume 3, Issue 10, October 2013.
15. M.Ramachandran "Relational Vs Non-Relationaldatabases", http://bigdatamadesimple.com/relational-vs-non-relational-databases, Accessed May 2015.
16. L.P.Issac, "SQL vs NoSQL Database Differences Explained with few Example DB".

A SURVEY ON APPLICATIONS OF IOT IN REAL LIFE SCENARIO

R.Kalaivani, M.Sc., M.Phil.,
Assistant Professor, Department of Information Technology
Madurai Sivakasi Nadars Pioneer Meenakshi Women's College, Poovanthi

Abstract
The current scenario of the world is full of data that is digital data. Anyone at anywhere can communicate, integrate, and collaborate with anybody, system and technology to go on with their work. This is due to the uprising in the field of Information and Communication Technologies (ICT) especially the evolution of IoT (Internet of Things). Today IoT touches every facet of our lives, opening new opportunities for growth, innovation and knowledge creation. It provides interconnection of people with many things (media, photos, information, etc.) and nowadays even with physical objects such as RFID, sensors, actuators, robots, etc.This paper provide a broad overview on what is IoTs, different applications of IOTs in our day to day life, and various challenges we are facing in implementation of this technology.
Keywords: *IoT, Applications, Security*

Introduction

The Internet of Things (IoTs) is a new paradigm which provides a large number of devices connected to the network, enabling "anytime, anywhere" access to information. It implies that these devices can be managed from the web and in turn; provide information in real time, allowing the interaction with people who use it (Gomez et al., 2013). The IoT model is evolving to accommodate any object capable of interacting directly with its local neighbor. In this context, the Internet can be viewed as a backbone network that interconnects a huge number of smaller (peripheral) networks, each of which would regroup objects according to its neighborhood relationships and physical properties. Examples of such smaller networks include sensor networks, vehicular networks and Mobile Adhoc Networks (MANETs) in general. The IoT will have a tremendous effect on all aspects of everyday life, promising to eventually provide identification, tracking and communication abilities to virtually every object on the planet [1]. The IoT will revolutionize networking over a myriad of applications, including participatory sensing, enhanced learning, e-health and automotive applications. Similarly, IoT's influence will reform numerous business disciplines such as intelligent manufacturing, retail, supply chains and product lifecycle management, in addition to reliable and safe transportation of people and goods [2], [3], [4], [5], [6],[7].

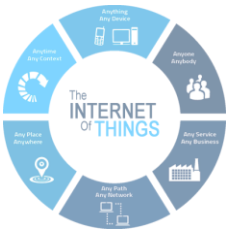


Figure 1 Internet of Things

Architecture of Iot

Internet of Things or IoT is an architecture that comprises specialized hardware boards, Software systems, web APIs, protocols which together creates a seamless environment which allows smart embedded devices to be connected to internet such that sensory data can be accessed and control system can be triggered over internet. Also devices could be connected to internet using various means like Wi-Fi,

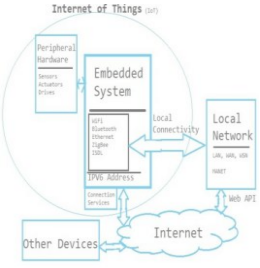


Figure 1: Architecture of IoT

Ethernet and so on. Furthermore devices may not need to be connected to internet independently. Rather a cluster of devices could be created and the base station or the cluster head could be connected to internet. This leads to more abstract architecture for communication protocols which ranges from high level to low level [8].

Applications of IoT

IoT applications are numerous and diverse, permeating into practically all areas of every-day life of individuals, enterprises, and society as a whole.

The 2010 Internet of Things Strategic Research Agenda (SRA) has identified and described the main Internet of Things applications, which span numerous applications — that can be often referred to as "vertical" — domains: smart energy, smart health, smart buildings, smart transport, smart living and smart city. The vision of a pervasive IoT requires the integration of the various vertical domains (mentioned before) into a single, unified, horizontal domain which is often referred to as smart life. The IoT application domains identified by IERC are based on inputs from experts, surveys and reports. The IoT application covers "smart" environments/spaces in domains such as: Transportation, Building, City, Lifestyle, Retail, Agriculture, and Factory, Supply chain, Emergency, Health care, User interaction, Culture and Tourism, Environment and Energy [8].

Figure 1: Applications of IoT

Literature Survey

In 2014, Partha P. Ray proposed a paper "Home Health Hub Internet of Things (H3IoT): An Architectural Framework for Monitoring Health of Elderly People". In this paper, he designed a framework which monitors the health of elderly people at home based on IoT [9].

In 2015, Bagheri, Maryam and Haghighi Movahed, Siavosh proposed a paper on "The effect of the internet of things (iot) on education business model".In this paper, they discussed about how IoT can shape smart campuses and classrooms.

They state that conversion of conventional classrooms into smart classroom will provide the advantage such as cost reduction, personalized learning, time saving, enhanced safety, improved comfort and collaboration [10]. In 2017, Dr. Vidya Gavekar and Dr. Manisha Kumbhar proposed a paper on "Role of Internet of Things for Integrating Smart Environment". Here they discussed about various applications of IoT and they provide a model smart city in which various departments are connected through IoT. The interconnected departments provide better solution and it reduces the time to tackle and solve a problem [8].

In 2017, Shwetal S. Raipure Rahul S. Khokale proposed a paper "Internet of Things for Automatic Traffic Control for Smart City".In this paper,they used LEACH algorithm to automatically control traffic system in smart city[11].

In 2017, Deepak Sharma proposed a paper on "Implementing Home Automation System using IoT for Electronic Devices". In this paper, He discussed about Home automation using IOT with integration of power level energy system. He proposed a model to connect all home appliances and control the usage of energy [12].

Challenges and Open Issues

The workflows in analyzed enterprise environment, home, office and other smart spaces in the future will be characterized by cross organization interaction, requiring the operation of highly dynamic and ad-hoc relationships.

The Key Challenges are

1. Network Foundation - limitations of the current Internet architecture in terms of mobility, availability, manageability and scalability are some of the major barriers to IoT.
2. Security, Privacy and Trust - in the domain of security the challenges are:
3. Securing the architecture of IOT - security to be ensured at design time and execution time.
4. Proactive identification and protection of IOT from arbitrary attacks (e.g. DoS and DDoS attacks) and abuse.
5. Proactive identification and protection of IOT from malicious software.

In the Domain of User Privacy, the Specific Challenges are:

1. Control over personal information (data privacy) and control over individual's physical location and movement (location privacy)
2. Need for privacy enhancement technologies and relevant protection laws
3. Standards, methodologies and tools for identity management of users and objects.

In the Domain of Trust, Some of the Specific Challenges are:

1. Need for easy and natural exchange of critical, protected and sensitive data - e.g. smart objects will communicate on behalf of users / organizations with services they can trust,
2. Trust has to be a part of the design of IoT and must be built in.

Managing heterogeneity - managing heterogeneous applications, environments and devices constitute a major challenge.

In addition to the above major challenges, some of the other challenges are:

1. Managing large amount of information and mining large volume of data to provide useful services
2. Designing an efficient architecture for sensor networking and storage
3. Designing mechanisms for sensor data discovery
4. Designing sensor data communication protocols - senor data query, publish/subscribe mechanisms
5. Developing sensor data stream processing mechanisms
6. Sensor data mining - correlation, aggregation filtering techniques design.

Finally, standardizing heterogeneous technologies, devices, application interfaces etc. will also be a major challenge [13].

Conclusion

This paper reviewed the divergent application of IoT and various researches done over this emerging technology. As IoT is going to occupy every sphere of human life, researches should give more concentration in the area of security and seamless integration of various technologies such as wireless communication, machine learning, embedded systems, wireless sensor networks and etc.

References

1. Atzori, L., Iera, A., Morabito, G. "The Internet of Things: A survey". Computer Networks, 54(15), 2787-2805, 2010.
2. G. Santucci, "The Internet of Things," in Between the Revolution of the Internet and the Metamorphosis of Objects", H. Sundmaeker, P.Guillemin, P. Friess, & S. Woelffle, (Eds.)Forum American Bar Association, pp. 11-24, Feb. 2010.
3. D. Yang, F. Liu, and Y. Liang, "A Survey of the Internet of Things,"in ICEBI-10,Advances in Intillegant Systems Research, ISBN 978-90-78677-40-6, Atlantic Press,December 2010.
4. L. Atzori, A. Iera, and G. Morabito, "The Internet of Things: A Survey," Int. Journal of Computer and Computer Networks, vol. 54, no. 15, pp. 2787–2805, Oct. 2010.
5. H. Sundmaeker, P. Guillemin, and P. Friess. "Vision and Challenges for Realising the Internet of Things". European Commission, ISBN:978-92-79-15088-3, March 2010.
6. D. Giusto, A. Iera, G. Morabito, and L. Atzori,"The Internet of Things", Heidelberg,Springer, 2010.
7. M. Zorzi, A. Gluhak, S. Lange, and A. Bassi, "From Today's INTRAnet of Things to aFuture INTERnet of Things: A Wireless and Mobility-Related View," IEEE Wireless Commun. , vol. 17, no. 6, pp.44-51, 2010.
8. Dr. Vidya Gavekar, Dr. Manisha Kumbhar, "Role of Internet of Things for Integrating Smart Environment", IJSRD - International Journal for Scientific Research & Development| Vol. 5, Issue 09, 2017 | ISSN (online): 2321-0613.
9. Partha P. Ray," Home Health Hub Internet of Things (H3IoT): An Architectural Framework for Monitoring Health of Elderly People", International Conference on Science, Engineering and Management Research (ICSEMR 2014), 978-1-4799-7613-3/14/ ©2014 IEEE
10. Bagheri, Maryam and Haghighi movahed, Siavosh, The Effect of the Internet of Things (IoT) on Education Business Model, http://shura.shu.ac.Uk /14405/
11. Shwetal S. Raipure, Rahul S. Khokale,"Internet of Things for Automatic Traffic Control for Smart City", IJSRD - International Journal for Scientific Research & Development| Vol. 5, Issue 09, 2017 | ISSN (online): 2321-0613
12. Deepak Sharma," Implementing Home Automation System using IoT for Electronic Devices", IJSRD - International Journal for Scientific Research & Development| Vol. 5, Issue 09, 2017 | ISSN (online): 2321-0613
13. Debasis Bandyopadhyay · Jaydip Sen," Internet of Things - Applications and Challenges in Technology and Standardization", Article in Wireless Personal Communications - May 2011 DOI:10.1007/s11277-011-0288-5, Source: arXiv

A COMPREHENSIVE SURVEY ON CLASSIFICATION TECHNIQUE IN BIG DATA

P.Surya
Assistant Professor, Department of Information Technology
Madurai Sivakasi Nadars Pioneer Meenakshi Women's College, Poovanthi, Tamil Nadu, India

Abstract
Big data is the growth in the volume of structured and unstructured data. Big data challenges include capturing data, data analysis, search, sharing, transfer, visualization, querying, updating and information privacy. Traditional technologies were not able to handle storage and processing of huge data thus Big Data concept comes into existence. This is a tedious job for users to identify accurate data from huge unstructured data. So, there should be some mechanism which classifies unstructured data into organized form which helps user to easily access required data. Classification techniques over big transactional database provide required data to the users from large datasets more simple way. There are many types of classification algorithm used in big data mining. In this paper mainly focus on four classification algorithm such as SVM, KNN.
Keywords: *Big Data, SVM, KNN.*

Introduction

Classification is one of the most useful techniques in data mining that classify data into structured class or groups. It helps the user in discovering the knowledge and future plan. Classification supplies the user with an intelligent decision making [1].

Big Data is a heterogeneous mix of data both structured (traditional datasets –in rows and columns like DBMS tables, CSV's and XLS's) and unstructured data like e-mail attachments, manuals, images, PDF documents, medical records such as x-rays, ECG and MRI images, forms, rich media like graphics, video and audio, contacts, forms and documents.

Classification technique is used to solve the many challenges which classify the big data according to the format of the data that must be processed, the type of analysis to be applied, the processing techniques at work, and the data sources for the data that the target system is required to acquire, load, process, analyze and store [2].

Big data analytics is the area where advanced analytic techniques operate on big data sets. It is really about two things, Big data and Analytics and how the two have teamed up to create one of the most profound trends in business intelligence (BI) [2].

Map Reduce by itself is capable for analyzing large distributed data sets; but due to the heterogeneity, velocity and volume of Big Data, it is a challenge for traditional data analysis and management tools [2].

A problem with Big Data is that they use NoSQL and has no Data Description Language (DDL) and it supports transaction processing. Also, web-scale data is not universal and it is heterogeneous.

Classification Technique Overview

Classification is one of the most useful techniques in data mining that classify data into structured class or groups. It helps the user in discovering the knowledge and future plan. Classification supplies the user with an intelligent decision making. Classification consists of two phases; the first phase is learning process phase in which a huge training data sets (BigData) are analyzed and then it create the patterns and the rules [3]. The second phase is evaluation or testing and recording the accuracy of the

performance of classification patterns. The purpose of classification is to be able to use its model to predict the class label of objects whose class label is unknown. Various forms can be represented for instance Neural Networks, classification rules, mathematical formulas or decision tree [4].In this paper discussed below two classification algorithm such as SVM and KNN.

Support Vector Machine (SVM)

Support Vector Machine (SVM) is the classification technique which used to process on large training data. The Big and complex data can be left to the SVM since the result of SVM will be greatly influenced when there is too much noise in the datasets. SVM provides with an optimized algorithm to solve the problem of over fitting. SVM is an effective classification model is useful to handle those complex data. SVM can make use of certain kernels to reveal efficiently in quantum form the largest eigenvalues and corresponding eigenvectors of the training data overlap (kernel) and covariance matrices [5].

Multilevel Clustering Algorithm Based SVM achieve on numerical data with the learning and evaluation of clustering algorithms. Achieve dimensionality lessening on the data to diminish noise, dimensions for correct use of it. So that is used for big data analysis [2]

Applications of SVM

SVM has been found to be successful when used for pattern classification problems. Applying the Support Vector approach to a particular practical problem involves resolving a number of questions based on the problem definition and the design involved with it. One of the major challenges is that of choosing an appropriate kernel for the given application [9].

There are standard choices such as a Gaussian or polynomial kernel that are the default options, but if these prove ineffective or if the inputs are discrete structures more elaborate kernels will be needed. By implicitly defining a feature space, the kernel provides the description language used by the machine for viewing the data. Once the choice of kernel and optimization criterion has been made the key components of the system are in place [10].

K-Nearest Neighbor (KNN)

The K-Nearest Neighbor classifier is one of the most well-known methods in data mining because of its effectiveness and simplicity. However, It lacks the scalability to manage big datasets. The main problems found for dealing with large-scale data are runtime and memory consumption [6].

The two approaches are developed under MapReduce paradigm. MR-KNN [6] approach developed in Apache Hadoop. This model allows us to classify large amounts of test examples against a big training dataset. The map phase computes the k-nearest neighbors in different splits of the training data [6].

Afterward, the reduce stage will compute the definitive neighbors from the list obtained in the map phase. Moreover, this parallel implementation provides the exact classification rate as the original k-NN model. The experiments show the promising scalability capabilities of the proposed approach [7] [8].

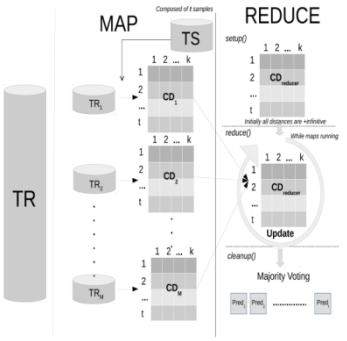


Figure 1: MR-KNN Classification

Applications of KNN

KNN as a data mining technique has a wide variety of applications in classification.

The KNN algorithm has been applied for analyzing micro-array gene expression data, where the KNN algorithm has been coupled with genetic algorithms, which are used as a search tool. Other applications include the prediction of solvent accessibility in protein molecules, the detection of intrusions in computer systems, and the management of databases of moving objects such as computer with wireless connections [11].

Conclusion

In this paper, we saw the different classification techniques on the era of Big Data. SVM and KNN techniques are better suited than the other for different applications. Each technique has a different accuracy, speed and predictors. K-Nearest neighbor technique to deal with large-scale problems and SVM is an effective classification model is useful to handle complex data.

References

1. A Survey of Classification Techniques in the Area of Big Data. PrafulKoturwar, SheetalGirase, Debajyoti Mukhopadhyay.
2. M. Minelli, M. Chambers, and A. Dhiraj, Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses (Wiley CIO), 1st ed. Wiley Publishing, 2013.
3. Mohammed GH. AL Zamil, "The Application of Semantic-based Classification on Big Data," International Conference on Information and Communication Systems (ICICS) 978-1-4799-3023 4/14, 2014, IEEE.
4. G. Kesavaraj, Dr. S. Sukumaran, "A Study on Classification Techniques in Data Mining," 4th ICCCNT – Tiruchengode, India, 31661, July 4 - 6, 2013, IEEE.
5. Dingxian Wang, Xiao Liu, Mengdi Wang, "A DT-SVM Strategy for Stock Futures Prediction with Big Data,"16th International Conference on Computational Science and Engineering 978-0-76955096-1/13, 2013, IEEE.
6. A MapReduce-based k-Nearest Neighbor Approach for Big Data Classification Jesus Maillo, Isaac Triguero, Francisco Herrera.
7. C. Zhang, F. Li, and J. Jestes, "Efficient parallel knn joins for large data in mapreduce," in Proceedings of the 15th International Conference on Extending Database Technology, ser. EDBT '12. New York, NY, USA: ACM, 2012
8. A. H. Project, "Apache hadoop," 2015. [Online]. Available: http: //hadoop.apache.org/
9. Burges C., "A tutorial on support vector machines for pattern recognition", In "Data Mining and Knowledge Discovery". Kluwer Academic Publishers, Boston, 1998, (Volume 2).
10. Nello Cristianini and John Shawe-Taylor, "An Introduction to Support Vector Machines and Other Kernel-based Learning Methods", Cambridge University Press, 2000.
11. Sadegh Bafandeh Imandoust And Mohammad Bolandraftar,"Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events: Theoretical Background".

A REVIEW ON KEY CHALLENGES IN INTEGRATION OF CLOUD COMPUTING AND INTERNET OF THINGS (IOT)

K.Sankareswari

Assistant Professor, Department of Information Technology

Madurai Sivakasi Nadar's Pioneer Meenakshi Women's College, Poovanthi, Tamil Nadu, India

Abstract

*Cloud computing and Internet of Things (IoT) are two different technologies and emerging technologies in modern internet era. But that are both already part of our life. IoT is upcoming emerging field limited with computational and storage capacity whereas cloud computing provides unlimited storage and computational capacity. Cloud computing is receiving a great deal of attention. Over the past several years, many companies have benefited from the implementation of cloud solutions within the organization. Due to the advantages such as flexibility, mobility, and costs saving, the number of cloud users is expected to grow rapidly. Most of the papers proposed the model for cloud and IoT separately. By combining cloud computing and Internet of Things together have lot of scope for research. In this paper mainly focus on the integration of Cloud and IoT(*Cloud based IoT *paradigm) and discuss about key challenges in integrating cloud computing and Internet of Things(IoT).*

Keywords: *Internet of Things Security, Cloud computing, Internet of things applications,* Cloud based IoT *paradigm*

Cloud Computing Basics

The term Cloud refers to a Network or Internet. In other words, we can say that Cloud is something, which is present at remote location. Cloud can provide services over network, i.e., on public networks or on private networks, i.e., WAN, LAN or VPN. Applications such as e-mail, web conferencing, customer relationship management (CRM), all run in cloud. Cloud Computing refers to manipulating, configuring, and accessing the applications online. It offers online data storage, infrastructure and application. Robin Hasting[1] has defined the cloud as "a massive network of cloud storage devices (servers and others) that exists somewhere over the internet. Cloud is a network of servers which is capable of running a service, providing storage opportunities or a platform to deliver certain tasks. Organization can adopt the cloud computing solution by outsourcing or setting up their own.

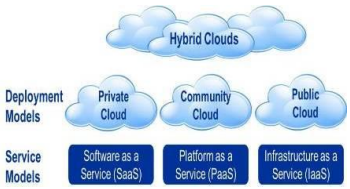
Cloud Computing Models

Cloud Computing Models[2] Cloud Providers offer services that can be grouped into three categories.

1. Software as a Service (SaaS): In this model, a complete application is offered to the customer, as a service on demand. A single instance of the service runs on the cloud & multiple end users are serviced. On the customers‟ side, there is no need for upfront investment in servers or software licenses, while for the provider, the costs are lowered, since only a single application needs to be hosted & maintained. Today SaaS is offered by companies such as Google, Salesforce, Microsoft, Zoho, etc.

2. Platform as a Service (Paas): Here, a layer of software, or development environment is encapsulated & offered as a service, upon which other higher levels of service can be built. The customer has the freedom to build his own applications, which run on the provider‟s infrastructure. To meet manageability and scalability requirements of the applications, PaaS providers offer a predefined combination of OS and application servers, such as LAMP platform (Linux, Apache, MySql and PHP), restricted J2EE, Ruby etc. Google‟s App Engine, Force.com, etc are some of the popular PaaS examples.

3. Infrastructure as a Service (Iaas): IaaS provides basic storage and computing capabilities as standardized services over the network. Servers, storage systems, networking equipment, data centre space etc. are pooled and made available to handle workloads. The customer would typically deploy his own software on the infrastructure. Some common examples are Amazon, GoGrid, 3 Tera, etc.



Types of Clouds

There are different types of clouds that you can subscribe to depending on your needs. As a home user or small business owner, you will most likely use public cloud services [11].

1. Public Cloud - A public cloud can be accessed by any subscriber with an internet connection and access to the cloud space.

2. Private Cloud - A private cloud is established for a specific group or organization and limits access to just that group.

3. Community Cloud - A community cloud is shared among two or more organizations that have similar cloud requirements.

4. Hybrid Cloud - A hybrid cloud is essentially a combination of at least two clouds, where the clouds included are a mixture of public, private, or community.

Internet of Things Basics

The term Internet of Things (often abbreviated IoT) [2] was coined by industry researchers but has emerged into mainstream public view only more recently. IoT is a network of physical devices, including things like smart phones, vehicles, home appliances, and more, that connect to exchange data with computers. The Internet of Things (IoT), sometimes referred to as the Internet of Objects, will change everything including ourselves. The Internet has an impact on education, communication, business, science, government, and humanity [1]. Clearly, the Internet is one of the most important and powerful creations in all of human history and now with the concept of the internet of things, internet becomes more favorable to have a smart life in every aspects [4]. Internet of Things is a new technology of the Internet accessing. By the Internet of Things, objects recognize themselves and obtain intelligence behavior by making or enabling related decisions thinks to the fact that they can communicate information about themselves [5]. These objects can access information that has been aggregated by other things, or they can add to other services [5].

Internet of Things Applications

Internet of things promises many applications in human life, making life easier, safe and smart. There are many applications such as smart cities, homes, transportation, energy and smart environment.

A. Smart Cities - Smart cities may still be viewed as a cities of the future and smart life, and by the innovation rate of creating smart cities today's, it will became very feasible to enter the IoT technology in cities development[6]. By the IoT, cities can be improved in many levels, by improving infrastructure, enhancing public transportation reducing traffic congestion, and keeping citizens safe, healthy and more engaged in the community.

B. Smart Home and Buildings - A smart home is a residence that uses internet-connected devices to enable the remote monitoring and management of appliances and systems, such as lighting and heating.Smart home technology, also known as home automation, provides homeowners security, comfort, Convenience and energy efficiency by allowing them to control smart devices, often by a smart home app on their smart phone or other networked device.

C. Smart Health - A close attention that required to hospitalized patients whose physiological status should be monitored continuously can be constantly done by using IoT monitoring technologies. For smart health sensors are used to collect comprehensive physiological information and use gateways and the cloud to analyze and store the information and then send the analyzed data wirelessly to caregivers for further analysis and review.

D. Smart Transportation and Mobility - The development in transportation is one of the factors to indicate the wellbeing of the country. A road condition monitoring and alert application is one of the most important of IoT transformation application [7]. The main idea of the concept of smart transportation and mobility is to apply the principles of crowd sourcing and participatory sensing.

E. Smart Factory and Smart Manufacturing - The smart factory will fundamentally change how products are invented, manufactured and shipped. At the same time it will improve worker safety and protect the environment by enabling low emissions and low incident manufacturing.

Combination of Cloud Computing and IOT

The combination of Cloud Computing and IoT can enable ubiquitous sensing services and powerful processing of sensing data streams beyond the capability of individual things, thus stimulating innovations in both fields. For example, cloud platforms allow the sensing data to be stored and used intelligently for smart monitoring and actuation with the smart devices. Novel data fusion algorithms, machine learning methods, and artificial intelligence techniques can be implemented and run centralized or distributed on the cloud to achieve automated decision making. These will boost the development of new applications, such as smart cities, grids, and transportation systems.

Issues Challenges in Integrating Cloud Computing and IOT

New challenges, however, arise when IoT meets cloud. Alessio Botta, Walter de Donato et al.,[9] authors proposed analyzing and discussing the need for integration, how these issues been tackled in literature.

1. Storage Resources: IoT involves by definition a large amount of information source, which produce a huge amount of non-structured or semi-structured like data volume (data size), variety (data types), and velocity (data generation, frequency). It implies collecting, accessing, processing, visualizing, archiving, sharing and searching large amount of data. Offering virtually unlimited, low-cost, and on demand storage capacity, Cloud is the most convenient and cost effective solution to deal with data produced by IoT.
2. Computational Resources: IoT device have limited processing resources that do not allow on-site data processing. Data collected is usually transmitted to more powerful nodes where aggregation and processing is possible. But scalability is challenging to achieve without a proper infrastructure. The unlimited processing capabilities of cloud and its on-demand model allow IoT processing needs to be properly satisfied and enable analyses of unprecedented complexity. Data-driven decision making and prediction algorithms would be possible at low cost and would provide increasing revenues and reduced risks. other perspectives would be to perform real-time processing, to

implement scalable, real time, collaborative, sensor-centric applications, to manage complex events and to implement task offloading for energy saving.

3. Communications: One of the requirements of IoT is to make IP-enabled devices communicate through dedicated hardware, and the support for such communication can be very expensive. Cloud offers an effective and cheap solution to connect, track and manage anything from anywhere at any time using customized portals and build-in apps.
4. New capabilities: IoT is characterized by a very high heterogeneity of devices, technologies, and protocols. Therefore, scalability, interoperability, efficiency, availability, and security can be very difficult to obtain. The integration with the cloud solves most of these problems also providing additional features such as case-of access, ease-of-use, and reduced deployment costs.

Conclusion

IoT devices can perform more efficiently by using the services of Cloud and storing and analytical process of data which becomes very effective and powerful when using cloud technologies. Out of these models security features are lacking more when compare to the performance. The main motivation behind is to provide a detailed study about the integration Cloud Computing and Internet of Things with perspective of applications, issues and challenges in IoT.

References

1. Tan, Lu, and Neng Wang. "Future internet: The internet of things." 2010 3rd International Conference on Advanced Computer Theory and Engineering (ICACTE). Vol. 5. IEEE, 2010.
2. Pradip Patil, Sumit Sharma, R.B. Gajbhiye, "A Study- Impact of Internet of Things (IOT) For Providing Services for Smart City Development", International Journal of Advance Research in Computer Science and Management Studies, Volume 3, Issue 6, June 2015.
3. M. A. Ezechina, K. K. Okwara, C. A. U. Ugboaja. The Internet of Things (Iot): A Scalable Approach to Connecting Everything. The International Journal of Engineering and Science 4(1) (2015) 09-12.
4. http://www.meraevents.com/event iotworkshop
5. http://www.nxp.com/assets/documents/data/en/white-papers/INTOTHNGSWP.pdf
6. https://www.thingworx.com/ecosystem/markets/smart-connected-systems/smart-cities/
7. Vahid Mirzabeiki. An Overview of the Freight Intelligent Transportation Systems; Division of Logistics and Transportation, Chalmers University of Technology.
8. Hong, Hua-Jun, et al. "Optimizing Cloud-Based Video Crowdsensing." IEEE Internet of Things Journal 3.3 (2016): 299-313.
9. Botta Alessio, Walter De Donato, Valerio Persico, and Antonio Pescapé. "On the integration of cloud computing and internet of things." In Future Internet of Things and Cloud (FiCloud), 2014 International Conference on, pp. 23-30. IEEE, 2014.
10. Nastic, Stefan, et al. "PatRICIA--A Novel Programming Model for IoT Applications on Cloud Platforms." 2013 IEEE 6th International Conference on Service-Oriented Computing and Applications. IEEE, 2013.
11. Stefan Nastic,Sanjin Sehie et al., international conference on Future Internet of Things and Cloud. 978-1-4799-4357-9/14, IEEE, 2014.
12. Abdelwahab, S.H.E.R.I.F., et al. "Cloud of Things for Sensing-as-a-Service: Architecture, Algorithms, and Use Case." (2012).
13. Kumrai, Teerawat, et al. "Multi-objective Optimization in Cloud Brokering Systems for Connected Internet of Things." (2012).

RGB Color Image Enhancement

R.Sheeba

Assistant Professor, Department of Computer Science,
Madurai Sivakasi Nadars Pioneer Meenakshi Women's College, Poovanthi, Tamil Nadu, India

Abstract

The major objective of image enhancement is to process a given low contrast image or blurred color image so that the result is clearer than the original image for a specific application. Image enhancement is used to enhance the equality and visual appearance of an image, or to provide a better transform representation for future automated image processing. Many images like medical images, satellite, aerial images and also real life photographs suffer from poor and bad contrast and noise. It is necessary to enhance the contrast and remove the noise to increase image quality. In this paper, color image enhancement using histogram equalization is proposed. The performance of proposed method is analyzed and compared with the results of multi scale retinex method. Performance is tested using some of the measurement techniques and elapsed time of execution.

Keywords: *Image enhacement; Color image; Histogram equalization; Enhancing color image*

Introduction

In digital image processing image enhancement is one of the preprocess step to improve the interpretability or perception of information in images for human viewers, or to provide better input for other automated image processing techniques. Color images provide more and richer information for visual perception than that of the gray images. Color image enhancement plays an important role in Digital Image Processing [1]. For improving the quality of the image and to give better input for processing the image, we use image enhancement technique [2]. In the proposed work the simulation model has been developed using MATLAB (R2013a) to study the effect by histogram equalization. The main objective of proposed work is the histogram equalization technique is used for enhance the color and contrast of a given image.

Image Enhancement

Image Enhancement modifies the visual impact that the image has on the interpreter in a fashion that improves the information content in the given image.

- Contrast enhancement
- Intensity, hue, and saturation transformations
- Density slicing
- Edge enhancement
- Making digital mosaics
- Producing synthetic stereo images

Image enhancement can be defined as perception of information for humans to get and provide good quality of input. It helps in modifying image attributes so that it becomes more suitable. While processing, more than one attributes can be modified. Rather these choices of attributes and how we modify are particular for a given task [3]. There are many techniques in order to enhance an image. These methods can be classified as following in two categories:

1. Spatial Domain Methods
2. Frequency Domain Methods

In spatial domain techniques, we directly deal with the image pixels. The pixel values are manipulated to achieve desired enhancement. In frequency domain methods, the image is first transferred in to frequency domain. It means that, the Fourier Transform of the image is computed first. All the enhancement operations are performed on the Fourier transform of the image and then the Inverse Fourier transform is performed to get the resultant image. Image enhancement is applied in every field where images are ought to be understood and analyzed. For example, medical image analysis, analysis of images from satellites etc. In this section we briefly describe the various image enhancement techniques [4]. Image Enhancement basically includes noise reduction from the image. Noises like Gaussian, salt and pepper adding into the given images then using successive filtering process. The processed filter image can analyzed with noise added image by the histogram equalization, provides level of intensity and shows image histogram.

RGB Color Model

RGB is an additive Color model represented with three primary colors, in which Red, Green and Blue light waves are added together to reproduce a broad array of colors. RGB is device dependent and quality of the white color depends on the nature of primary light sources. Its Color Components are red, blue and green each has value in the range 0-255 [6].

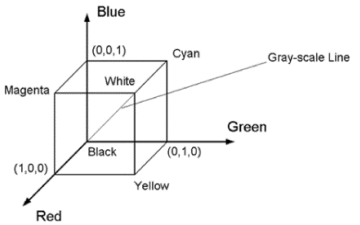


Figure 1: Flow chart of RGB color image enhancement process

Histogram Equalization

Histogram equalization is to improve the contrast of an image and transform an image in such a way that the transformed image has a nearly uniform distribution of pixel values [1].Histogram equalization is a common technique for enhancing the appearance of images. Presume we have an image which is mostly dark. Then its histogram would be skewed towards the lower end of the grey scale and the entire image feature is solid into the dark end of the histogram. If we could `stretch out' the grey levels at the dark end to produce a more uniformly distributed histogram then the image would become much clearer [7].

Transformation

Assume r has been normalized to the interval [0, 1], with r = 0 representing black and r = 1 representing white.

s=T(r) 0<r <1

The transformation function satisfies the following conditions:

- $T(r)$ is single-valued and monotonically increasing in the interval 0 < r <1
- 0 < T(r) <1 for 0 < r <1

Histogram equalization is based on a transformation of the probability density function of a random variable.

Let $p_r(r)$ and $p_s(s)$ denote the probability density function of random variable r and s, respectively. If $p_r(r)$ and $T(r)$ are known, then the probability density function $p_s(s)$ of the transformed variable s can be obtained

$$p_s(S) = p_r(r)\left|\frac{dr}{ds}\right|$$

Define a transformation function $s = T(r) = \int_0^r p_r(w)dw$

Where w is a dummy variable of integration and the right side of this equation is the cumulative distribution function of random variable r.

Given transformation function $T(r), T(r) = \int_0^r p_r(w)dw$

$$\frac{dr}{ds} = \frac{dT(r)}{dr} = \frac{d}{dr}\left[\int_0^r p_r(w)dw\right] = p_r(r)$$

$$p_s(s) = p_r(r)\left|\frac{dr}{ds}\right| = p_r(r)\left|\frac{1}{p_r(r)}\right| = 1\ 0 \le s \le 1$$

$p_s(s)$ now is a uniform probability density function.

$T(r)$ depends on $p_r(r)$, but the resulting $p_s(s)$ always is uniform.

In discrete version

The probability of occurrence of gray level r_k in an image is

$$p_r(r) = \frac{n_k}{n}\ k = 1,2,\dots,L-1$$

n: the total number of pixels in the image

n_k: the number of pixels that have gray level r_k

L: the total number of possible gray levels in the image

The transformation function is

$$s_k = T(r_k) = \sum_{j=0}^{k} p_r(r_j) = \sum_{j=0}^{k} \frac{n_j}{n}\ k = 0,1,2,\dots,L-1$$

Thus, an output image is obtained by mapping each pixel with level r_k in the input image into a corresponding pixel with level s_k.

Proposed Work

The proposed method is designed to enhance the RGB color image. The current work implements a system for the improved enhancement of RGB color images using histogram equalization. Fig.2 shows the steps for the proposed method. The proposed technique has the ability to produce efficient results even in case of high density of the noise.

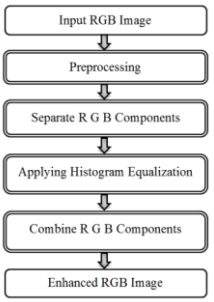


Figure 2: Flow chart of RGB Color image enhancement process

Preprocessing

The aim of pre-processing is an improvement of the image data that suppresses undesired distortions or enhances some image features relevant for further processing and analysis task.

Noise Reduction: Image noise is random variation of brightness or color information in image that is included due to environment and camera sensors. Gaussian noise and Salt-and-Pepper noise are most occurring noise in the images. It is impossible to eliminate noise completely, but noise can be suppressed using filters like Linear smoothing, Median [5].

Changing Sharpness or Brightness of image: Lightness or Darkness of image pixel can be referred as Brightness of image whereas Sharpness is defined as edge contrast. If brightness increases, lighting effect in the image increases. Similarly, if Sharpness increases smoothness of edges increases [5].

Histogram equalization: It is a method of adjusting contrast of image which usually increases global contrast that allows for areas of lower local contrast to gain higher contrast. This can be accomplished by effectively spreading out most frequent intensity values [5].

De-blurring: Image blur is difficult to avoid in many situations and can often ruin a photograph. Images blur causes decrease of quality in the image and left with unclear images. Smoothing filters are generally used for de-blurring images [5].

Process of Color Image Enhancement

- Preprocess the given RGB image.
- Separate R, G and B components from the given preprocessed image.
- Apply histogram equalization on each band.
- Combine R, G and B components as single image.

Results and Discussion

Fig. 3 (a) shows the noisy and low quality input images. Fig. 3 (b) shows the resultant images obtained after applying each step explained above. The various input images which are in low quality, blurred and enhanced images for those images are shown. The enhanced images in Fig. 3 are much clearer than input images. The multi scale retinex method [10] also applied on each input image to perform the comparison with the proposed method and their results are given in the below section.

(a) (b)

Figure 3 : (a) Original images; (b) Enhanced images

Performance Measurements

The performance results of proposed and multiscale retinex methods are given in this section. For measuring the performance of an enhanced image, two parameters are used i.e. entropy and contrast.

Entropy

Here, we use entropy as the Shannon entropy which contains the maximum information. It is used to measure the quality of image. Images which have maximum entropy have the better quality of the image [8]. Entropy differences between original and enhanced images are given in Table 1.

$$Ent(p) = -\sum_{I=0}^{L-1} p(l) log l$$

Where, Ent = entropy of the image; P (l) = probability density function at intensity level l; L = total no of gray level

S. No	Image Name	Entropy		
		Input Image	Histogram Equalization	Multiscale Retinex
1.	Image 01	7.0246	5.9286	7.5326
2.	Image 02	7.5079	5.9872	7.2544
3.	Image 03	7.4956	5.9842	7.5342
4.	Image 04	7.3141	5.9699	7.2723
5.	Image 05	7.6736	5.9755	7.6690

Table 1: Entropy Values of Enhanced Images

Contrast

Second parameter is contrast which is also used for measuring the quality of the image. Image which have highest value of contrast have good contrast visually [8]. Contrast differences between two images are given in Table 2.

$$\text{Contrast} = \sum_{i,j} |i - j|^2 \, p(i, j)$$

Where, p (i, j) = no of gray level occurrence

S. No	Image Name	Contrast		
		Input Image	Histogram equalization method	Multi scale retinex method
1.	Image 01	0.7147	0.9188	0.7137
2.	Image 02	0.7288	0.8664	0.6180
3.	Image 03	0.6220	0.9019	0.7320
4.	Image 04	0.6983	0.9639	0.7692
5.	Image 05	0.8933	0.9400	0.8573

Table 2 : Contrast values of Enhanced Images

Standard Deviation

It is the deviation about mean. It represents the dynamic range of values present in an image about the Mean. Standard deviation of each enhanced images are given in Table 3.

S. No	Image Name	Standard Deviation		
		Input Image	Histogram Equalization	Multiscale Retinex
1.	Image 01	43.0284	74.6971	50.4078
2.	Image 02	47.5637	74.8298	40.1858
3.	Image 03	56.6610	74.7418	56.1041
4.	Image 04	44.1478	74.7574	39.3893
5.	Image 05	61.8042	74.7016	62.9082

Table 3: Standard deviation of enhanced images

In order to know the efficiency of the proposed method some of the parameters like Peak Signal to Noise Ratio (PSNR), Mean Squared Error (MSE) have been used.

MSE

The mean squared error (MSE) of an estimator measures the average of the squares of the errors [9].

$$\mathrm{MSE} = \frac{1}{\mathrm{MN}} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} [X(i,j) - Y(i,j))]^2 \qquad (3)$$

PSNR

Peak signal-to-noise ratio (PSNR) is an expression for the ratio between the maximum possible value (power) of a signal and the power of distorting noise that affects the quality of its representation [9].

$$\mathrm{PSNR} = \frac{10 \log_{10} (L-1)^2}{\mathrm{MSE}} \qquad (4)$$

Now the low contrast and blurred images are enhanced using the proposed method. The PSNR and MSE values are calculated using by the equations (3) & (4) for the enhanced images and the values are given in Table IV. In general, PSNR value must increase and MSE value must decrease for a good segmented image. The PSNR and MSE values are compared between original images and the tumor detected images.

Image Name	MSE		PSNR	
	Histogram equalization	Multi scale retinex	Histogram equalization	Multi scale retinex
Image 01	2.9429e+03	3.0463e+03	30.9537	30.6084
Image 02	2.0343e+03	3.7904e+03	34.6461	28.4231
Image 03	2.9403e+03	1.3051e+03	30.9625	39.0852
Image 04	3.4657e+03	960.0307	29.3186	42.1556
Image 05	2.7356e+03	5.2862e+03	31.6844	25.0966

Table 4 : PSNR and MSR Values of Enhanced Images

Elapsed Time

Elapsed time is time taken to enhance the color image of the given input image. The execution time of this method is calculated by tic and toc methods using MATLAB. The elapsed time of different enhanced images is shown in Table 5.

Image Name	Elapsed Time (Seconds)	
	Histogram equalization	Multi scale retinex
Image 01	0.387675	6.464921
Image 02	0.345111	2.523471
Image 03	0.739701	28.383103
Image 04	0.343401	2.230775
Image 05	0.542073	10.036569

Table 5: Execution Time of the Proposed Work

Conclusion

Image enhancement processes offer a wide variety of approaches for modifying images to achieve visually acceptable images. In this paper, the low quality images are enhanced by using histogram equalization technique and enhanced. This technique is applied to 50 low quality images including low quality images, old photographs using MATLAB. Some of the resultant images are shown in Fig. 5(b). Histogram equalization method applied on every image in efficient manner. From the enhanced image it is possible to get the detailed information about that image with good visibility. It sharpens the image features such as edges, boundaries, or contrast to make a graphic display more helpful for display and analysis. It improves the clarity of images for human viewing, removing blurring and noise, increasing contrast, and revealing details. The efficiency of enhanced color image is measured by using contrast and entropy. Error rate is calculated by using PSNR and MSE methods. The results of proposed method are better than multiscale retinex method. This proposed method is simpler than other color image enhancement techniques and computationally fast.

Future Scope

In future this proposed color image enhancement technique can be used to segment the regions of an image based on the specific color. Also this technique can be improved to enhance the damaged color images.

References

1. Rafael, C. Gonzalez and Richard E. Woods. "Digital Image Processing", 2nd edition, Prentice Hall, 2002.
2. P. Janani, J. Premaladha and K. S. Ravichandran, "Image Enhancement Techniques: A Study", Indian Journal of Science and Technology, vol 8(22), DOI: 10.17485/ijst/2015/v8i22/79318, September 2015.
3. Raman Maini and Himanshu Aggarwal, "A Comprehensive Review of Image Enhancement Techniques", Journal Of Computing, Volume 2, Issue 3, ISSN: 2151-9617, March 2010.
4. Urvashi Manikpuri and Yojana Yadav, "Image Enhancement Through Logarithmic Transformation", International Journal of Innovative Research in Advanced Engineering (IJIRAE) Volume-1, Issue-8,ISSN: 2349-2163 pp. 357-362, September 2014.
5. Isunuri Bala Venkateswarlu, "Analytical Survey of Colour to Greyscale Conversion Methods Based on Primary Image Processing Techniques", International Journal of Advanced Research in Computer Science and Software Engineering, Volume - 4, Issue - 10, October 2014.
6. Haneet Kour, "Analysis on Image Color Model", International Journal of Advanced Research in Computer and Communication Engineering Volume - 4, Issue - 12, December 2015.
7. Kinita b vandara and G.r.kulkarni, "Spatial Domain Image Enhancement And Restoration Techniques", Journal of Information, Knowledge And Research In Electronics And Communication, Volume - 02, Issue -01, ISSN: 0975 – 6779, October 2012.
8. Kanika Kapoor and Shaveta Arora, "Colour Image Enhancement Based On Histogram Equalization", Electrical & Computer Engineering: An International Journal (ECIJ) Volume - 4, Number - 3, September 2015.
9. N.Jayachandra, SP.Valliammai, "Automated Diabetic Retinopathy Detection with Image Segmentation Techniques", International Journal of Digital Communication and Networks (IJDCN) Volume - 3, Issue - 3, March 2016.
10. Ana Belen Petro , Catalina Sbert and Jean-Michel Morel, "Multiscale Retinex", Image Processing On Line, ISSN 2105–1232,April 2014.

AUTOMATIC CERTIFICATE GENERATOR

S.Karpagaselvi
II M.Sc., (CS & IT), Department of Computer Science
Madurai Sivakasi Nadar's Pioneer Meenakshi Women's College, Poovanthi, Tamil Nadu, India

Abstract
The "Automatic Certificate Generator" has been developed on PHP and MYSQL. The main aim of this project is to develop a certificate generator which can be used in colleges; universities to manage all the events conducted and to automate the distribution of student certificates in portable document format (PDF).It can reduce the manual work of writing certificates. It also provides ready-to-use certificates which could be created and will be ready to print.
Keywords: *Certificate Generator, mpdf – class library*

Introduction

Automatic Certificate Generator is mainly used to create certificates systematically to avoid manual process of writing certificates. It is used in colleges, universities to manage all the events. The details of the students can be automatically filled and provides certificates in portable document format (PDF) which is globally accepted format for files and ready to print. The main feature of this system is to create bonafide certificate, certificates of events, facility to print and easily export PDF of the automatically generated certificates. It is flexible for generating the certificates of the students.

Existing System

In the existing system colleges and universities can create certificates manually as handwritten. It takes lot of efforts and also takes more time to write certificates with all the student details. The person who requires certificate has to wait for some time. Information accuracy on certificates can be low. Manual errors on student information can also occur during writing certificate.

Proposed System

Certificate generation system can automate the process of writing certificates in user friendly manner by not making it very complex. The system being automated and generalized, and ensures to reduce manual errors by reducing manual efforts. The student details can be filled automatically without any manual errors. It can be stored in the database so that we can refer it in future. Administrator can have the only rights to create certificates. For that login authentication can be done. The workload or manpower can be reduced in this system. The certificates can be created in just one click without any delay so the time taken for generating certificates can be reduced. We can also store the certificates for future references. It has the facility to export the certificates in PDF format files. We can also print certificates on the spot. Bonafide certificate, fees structure and certificates for all events conducted in our college can be generated.

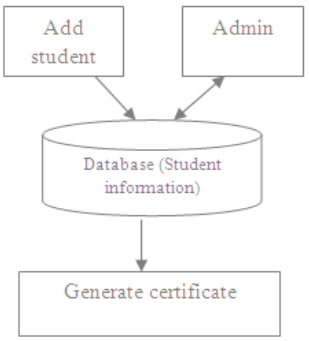


Figure 1: Architecture of the implemented system

Working of Certificate Generator

The administrator has the whole responsibility of student registration, updating the data from database, deleting the data, downloading certificates. In the Login page administrator can only login using the appropriate Username and password. The student registration can be done to store the details of student like name, department etc. in the database. Finally the system will generate the certificates as PDF file by fetching student details for printing purpose. MPDF class is used in the system to generate the certificate in PDF format. The system has been developed for our college (Madurai Sivakasi Nadars Pioneer Meenakshi Women's College) and implemented in an effort to make it as attractive and dynamic as possible.

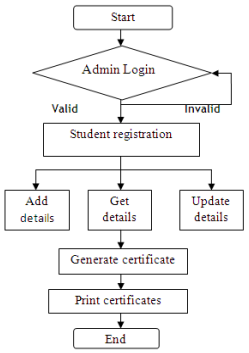


Figure 2 Flow chart of certificate generator

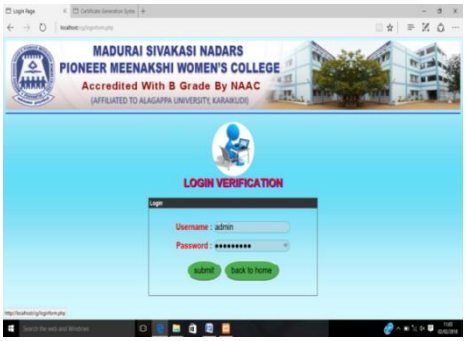


Figure 3: Screenshot of the login interface

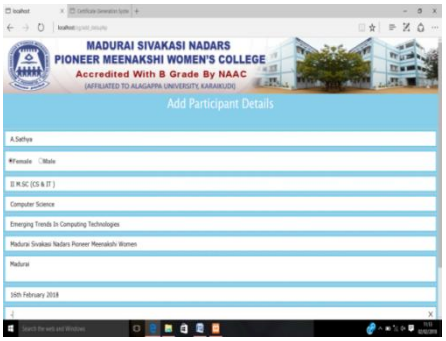


Figure 4: Screenshot of student registration

Figure 5: Screenshot of the generated certificate

MPDF Class Library

MPDF is a simple and popular tool for shared hosting users to create and convert UTF-8 encoded HTML pages to PDF files. MPDF was written by Lan Back. In other words MPDF is a PHP classes based on FPDF and HTML2PDF tools which purpose is the same – to create and manage PDF documents using simple PHP. By using MPDF we can create an HTML file structure which will be converted to PDF format file. We can also use some CSS coding and improve your files. However, there are some restrictions using HTML tables and styling text inside cells. With MPDF we will be able to create rich PDFs with headers and footers, CSS styling, included pictures, watermarks, pages numbering, and justification.

MPDF class library provides and support Unicode Characters and font, character substitution methods. Some of the features of MPDF class library are as follows:

- A single CSS style sheet can be used for all pages, with font substitution automatically for CJK characters.
- Character substitution can optionally be used to automatically replace any characters that do not exist in the current font.
- Word spacing and character spacing are both used to justify text; works in Unicode mode and CJK characters as well.
- Nested block-level elements (e.g. P, DIV) are supported. For e.g. margins, borders, padding, line-height, background colors etc.
- CSS style attributes fully support font, font-size, color, and background color, Table cell padding and borders are supported.
- Watermarks as text or images can be used and will appear as a transparency over other elements.
- Multiple style sheets can be added within a single page

References

1. Steven Holzner, "PHP The Complete Reference" Tata Mcgraw-Hill Edition 2008, ISBN-13:978-0-07-022362-2
2. Vikram Vaswani, "PHP Programming Solutions" Tata Mcgraw-Hill Edition 2008, ISBN-13:978-0-07-065978-8
3. MPDF manual, features, functions http://www.mPDF-Manual.com
4. "Certificate generation system" – International journal on recent and innovation trends in computing & communication,
5. http://www.IRJITCC.org/certificate_generation_system
6. Kevin yank, "Build your own database driven website using PHP & MYSQL" 4th edition, Site point publications, http://www.it-ebooks.info/php&mysql
7. "MYSQL tutorial- simple and easy learning", http://www.tutorialpoints.com/mysqltutorials
8. "Creating pdf file using mpdf & php" http://www.phpflow.com/php/create-pdf-file-using-mpdf-php/
9. Patrick T. Lane, "Advanced CSS & HTML5guide, http://ciwcertified.com/document /html5.pdf
10. Matt Doyle, "Beginning PHP5.3", http://cs.petrsu.ru/php/beginning-php5.3 by matt doyle.pdf
11. "MPDF documentation and manual", http://mpdf.github.io/mpdf-manual

TIMETABLE GENERATOR

M.Sonavarshini
II-M.Sc (CS & IT), Department of Computer Science,
Madurai Sivakasi Nadars Pioneer Meenakshi Women's College, Poovanthi, Tamil Nadu, India

Abstract
The Project entitled "Timetable Generator" has been developed on PHP and MYSQL. It is developed to provide better support for lecture and student in a college. It is an automated system which generate timetable according to the data given by the user. The main requirement of the application is to provide the details about the subject, staff details, workload. Then the application generates the timetable according to your need. It reduce the time requires to generate time table which is more accurate, precise and free of human errors. This project introduces a automated timetabling capable of taking care of both strong and weak constraints effectively. Timetable Generation System generates timetable for each class.
Keywords: *Timetable Generator*

Introduction

It is capable taking care of hard and soft constraints required specifically for preparing timetable in colleges. The purpose of this to built an application program to reduce the manual work for managing timetable. By automating this process of timetable generator can save a lot of precious time of administrators who are involved in creating and managing course timetables. Staff timetable and student timetable are generated successfully. The aim is here to develop a simple, easily understandable, efficient and portable application which could automatically generate good quality time table.

Existing System

The traditional hand operated method of timetable is very time consuming and usually ends up with various classes clashing either same teacher having more than one class at a time which is being resolved by automated timetable scheduling. The manual timetable scheduling demands considerable time and efforts along with lots of paperwork.

Proposed System

It introduces a Practical timetable approach capable taking care of both hard and soft constraints required specifically for preparing timetable in colleges. The automated timetable scheduling provide easier way for teacher and student to view timetable once they are finalized over the application having individual login id and password.

Constraints

No students can attend more than one classes, No teacher can teach more than one subject at a time.

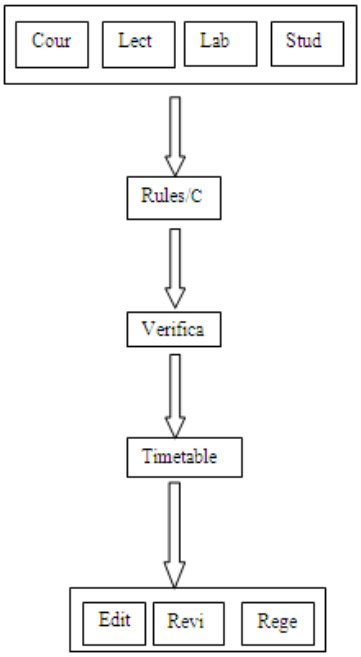


Figure 1: Flow Chart

About the Project

The Timetable Generator has been created for our college Madurai Sivakasi Nadars Pioneer Meenakshi Women's College. The Timetable Generator in which each staff has an individual login id and password. Subject, Staff, Course, Department details are manipulated such as Insert, Delete, Update, View. Session is created so far the information are stored until staff Logout. The Main Thing is the staff can select three subjects willing to take. As well as the same subject is not select more than thrice as the subject is hidden. The subjects are stored and according to the workload of the staff members the subject are divided and construct a timetable. Thus the timetable is generated successfully.

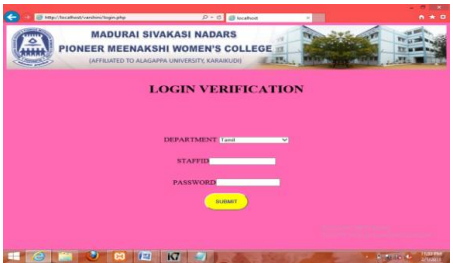
Implementation

There are two types of login, administrator login, Staff login,

Admin Login

The timetable generation system allows only admin to log in to the system means authentication is provided to admin only. Therefore entire system will be managed by admin.

Staff Login



The staff has each individual staff Id and password to login the page.

Admin Page

After successful log in of admin the page provides different option, each of this process hold the specific information such as update, delete subjects and lectures details which can be done by the admin. The admin if need then add department, staff, subjects, courses. Logout option will move admin to the home page.

Course Details

Subject Details

The subject details include Department Id, Course Id, Type, Subject Name, Subject Code. These are store in the database in which admin can manage the information.

ID	DEPARTMENT ID	COURSE ID	TYPE	SUBJECT NAME	SUBJECT CODE
1	6	2	Major	Programming in C&DataStructures	7BIT2C1
2	6	2	Major	DataStructuresUsingCLab	7BIT2P1
3	6	2	Major	Java Programming	4BIT4C1
4	6	2	Major	SoftwareEnginnering	4BIT6C1
5	6	2	Major	CloudComputing	4BIT6C2
6	6	2	Major	WebProgramming	4BIT6C3
7	6	2	Allied	DesktopPublishing	7BSOA2
8	6	2	Allied	ComputerOrientedNumericalMethods	4BCESA4
9	6	2	Allied	MobileCommunication	4BITE3A
10	6	2	Others	EnvironmentalStudies	7BES2

Staff Details

ID	STAFFID	DEPARTMENT	NAME	PASSWORD	DESIGNATION
1	msnMAT01	Mathematics	C.Ramalakshmi	MAT01	H.O.D
3	msnIT01	InformationTechnology	K.Mahalakshmi	IT01	H.O.D
5	msnCS01	ComputerScience	K.Sudharani	CS01	H.O.D
7	msnCA01	Commerce(CA)	N.Gomathi	CA01	H.O.D
9	msnIT06	Information Technology	K.Sankareswari	IT06	Asst.Professor
11	msnTAM01	Tamil	R.Poongkuzhali	TAM01	H.O.D
12	msnTAM02	Tamil	P.KarthigaiSelvi	TAM02	Asst.Professor
13	msnTAM03	Tamil	S.Kalaivani	TAM03	Asst.Professor
14	msnTAM04	Tamil	V.AntonyTamilarasi	TAM04	Asst.Professor
15	msnTAM05	Tamil	P.Seethalakshmi	TAM05	Asst.Professor

Conclusion

It is used to simplify manual work. It minimize document related work. Friendly environment. The future enhancement is to considering the room as an input and allocate the classes while generating the timetable. And additional features are that faculty replacement is also made possible by listing out the available faculty until a replacement faculty is assigned. This enhancement can be achieved my making further modification

References

1. Programming PHP - Rasmus Lerdorf
2. PHP and MySQL web development - LukeWelling
3. Web database Applications with PHP and MySQL - HughE.William
4. PHP Programming - MladenGogala
5. Learning PHP Design Patterns - Williams Sanders.
6. Web Technology L. Mathu Krithiga Venkatesh
7. The Complete Reference PHP - StevenHolzner
8. Sams Teach Yourself PHP, MySQL and Apache-Julie c. Meloni
9. PHP programming Solutions - Vikram Vaswani
10. Web Technology A developer Perspective-N.P. Gopalan, J. Akilandeswari

WEBSITE FOR THE DEPARTMENT

V.Jayalakshmi
II M. Sc. (CS & IT), Department of Computer Science
Madurai Sivakasi Nadars Pioneer Meenakshi Women's College, Poovanthi, Tamil Nadu, India

Abstract
The project entitled "Website for the Department" has been developed using php as Front End and MYSQL as Back End. The main objective of website is to automate all functionality of a department. In the college website the details of the department can only be viewed. The website for the department can be developed to post and editing all the details. To developed the website with dynamic and interactive.
Keywords*: Department website, Blog*

Introduction

A website is a collection of Web pages, images, videos that hosted on Web server and accessible via Internet connection. A website is a most powerful tool for communication. The website pages can access from common domain name and published on web server. Web pages, which are the building blocks of websites, are documents, typically composed in plain text with formatting instructions of Hypertext Markup Language (HTML). A website requires attractive design and proper arrangement of links and images, which enables a browser to easily access.

About the Project

The website has been created for our college (Madurai Sivakasi Nadars Pioneer Meenakshi Women's). The project starts with a home page containing a directory of the site of web content and can be used to display the vision, mission and objective of our department. The faculty details, publications are stored in database using sql statements such as insert, delete, update and view. Student syllabus can be post on the website. Events are sent through e-mail to other colleges. Our department contains several activities such as Association, Extension activities, Team teaching, Peer team teaching, Certificate course, intercollegiate competition, Bridge course and Industrial visit. The new events are marquee on the home page such as announcements and events associated with the department activities. Head of the department have the permission to post the events.

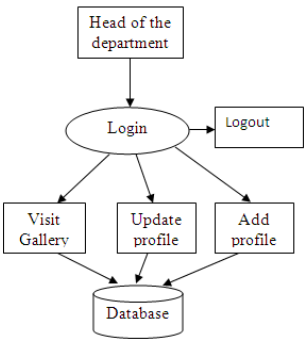


Figure 1: Flowchart for Admin

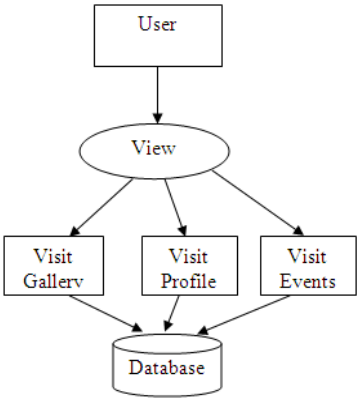


Figure 2: Flowchart for user

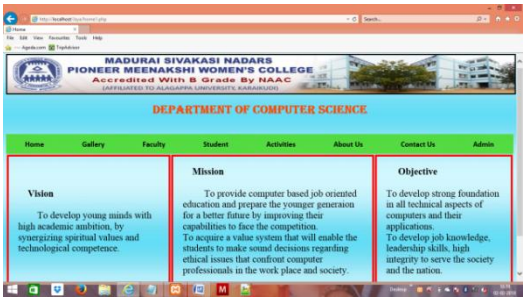


Figure 3: Screenshot of Home page

Figure 4: Screenshot of Event Gallery

Peer Team Teaching

A process between two or more students in a group where one of the students acts as a teach for the other group mates. It can be applied among students of the same age or students belonging to different age groups.

Types of Peer Team Teaching

1. Instructional peer teaching - older students help younger ones on a one-to-one or one-to-a group basis.
2. Same age peer teaching - more able ones to assists the less able.
3. Monitorial peer teaching - the class maybe divided into groups and monitors are assigned to lead each group.

Figure 5: Screenshot of Events

Team Teaching

A method of Co-ordinated classroom teaching involving a team of teachers working together with a single group of students.

Types of Team Teaching

1. A team of teachers from same department.
2. A team of teachers from inter-department but from the same institutions.
3. A team of teachers from inter-institution.

Bridge Course

A bridge course is a university preparation course with an academic curriculum that is offered to mature students as a means of preparing for the intellectual challenges of a university education, successful completion of which is recognized as a basis of admission to the university.

Association

A group of people in our college having a common purpose or interest. The act of association is to do increase our extracurricular activities. Then the students are participating in each and every activity.

Industrial Visit

An objective of industrial visit is to provide students on insight regarding internal working of companies. Industrial visit provide an excellent opportunity to interact with industries and know more about industrial environment. Industrial visit are arranged by colleges to students with an objective of providing student functional opportunity in different sectors like IT, Manufacturing and services, finance and marketing. It helps to combine theoretical knowledge with industrial knowledge. Industrial realities are opened to the students through industrial visit.

Extension Activity

Extension activities provide a link between the college and society. In order to create socially sensitive citizen, the students are made aware of the common extension activities through NSS, RRC, YRC and department specific extension activities during the induction programme.

Certificate Course

Certificate programs are commonly offered in both professional and academic subject areas. In our college each and every departments are offered certificate course. In our department we offered certificate course such as Flash, .Net and J2EE.

E-Mail

Electronic Mail (**email** or **e-mail**) is a method of exchanging messages between people using electronic devices. E-mail messages are usually encoded in ASCII text. Our department create own mail server to send the information such as Seminar invitations. Any new events can be conducted in our department that information can send to other colleges through email.

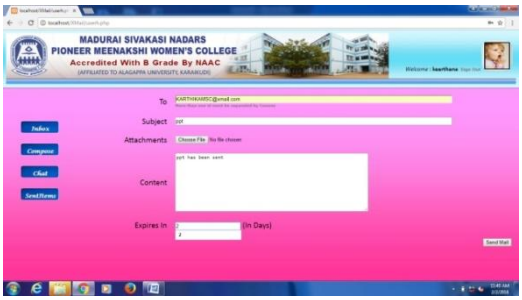


Figure 6: Screenshot of E-Mail

Blog

A blog is a type of website that is updated regularly with new content. Most blogs contain short, informal activities called blog posts. These posts usually contain some combination of text, image, video and other media. To express our opinions, experience and interest. The people who write blogs are called bloggers. Head of the department can insert the new post, update, delete, view the posts and responsibilities for update the database by adding new information's.

Conclusion

This project has come to final stage. The system has developed with free of errors, efficient and take less time consuming. It was wonderful and learning experience for me while working on this project. In this project I came to know how professional software's are designed. I enjoyed each and every bit of work in this project. The project has covered almost all the requirements. Further requirements and improvements can easily be done since the coding is mainly structured or modular. In other feature, we will add video as possible. I couldn't implement the features, so I have done in future. Further enhancements can be made to the department, so that the website functions very attractive. This Online system will be implemented soon.

References

1. Julie C. Meloni, "Sams Teach Yourself PHP, MYSQL and Apache", All in One, Fourth Edition, PEARSON.
2. Vikram Vaswani, "PHP PROGRAMMING SOLUTIONS" Tata McGraw Hill Edition 2007.
3. David Powers, "PHP Solutions: Dynamic Web Design Made Easy".
4. Robin Nixon, "Learning PHP, MySQL and Javascript".
5. Larry Ullman, "PHP for the world wide web".
6. Rasmus Lerdorf, "Programming PHP".
7. Mike McGrath, "PHP and MySQL".
8. George Schlossnagle, "Advanced PHP Programming".
9. www.phppoint.com
10. www.phpplot.com

DATA MINING: A REVIEW ON EDUCATION DATA MINING TECHNIQUES

N.Vinothini

Assistant Professor, Department of Information Technology

Madurai Sivakasi Nadars Pioneer Meenakshi Women's College, Poovanthi, Tamil Nadu, India

Abstract

Educational Data Mining (EDM) is an emerging multidisciplinary research area, in which methods and techniques for exploring data originating from various educational information systems have been developed. It provides intrinsic knowledge of teaching and learning process for effective education planning. In this survey work focuses on education data mining methods, phases, goals, issues and applications of education data mining.

Key Words: *Data Mining, Education data mining, offline data, online data, Classification, Clustering, outlier detection, student modelling and predictive modeling.*

Introduction

Data mining is the task of discovering interesting patterns from large amounts of data, where the data can be stored in databases, data warehouses, or other information repositories. It is a young interdisciplinary field, drawing from areas such as database systems, data warehousing, statistics, machine learning, data visualization, information retrieval, and high-performance computing. Other contributing areas include neural networks, pattern recognition, spatial data analysis, image databases, signal processing, and many application fields, such as business, economics, and bioinformatics [1].

Education Data Mining (EDM) is the application of data mining related to educational data and Educational Data Mining is a learning analytics and quantitative observation method in order to understand how student respond to educational system and their responses impact their learning. Its objective is to analyze educational data in order to resolve educational research issues. In recent years there is rapid growth in education sector which leads to growing of education data so mining of education data become important to understand student behavior during learning process or to understand student problems[2]

Types of Educational Data

Offline Data - Offline Data are generated from traditional and modern classroom, Interactive teaching/learning environments, learner/educators information, students attendance, Emotional data, Course information, data collected from the academic section of an institution etc..

Online Data - Online Data are generated from the geographically separated stake holder of the education, distance educations, web based education and computer supported collaborative learning used in social networking sites and online group forum [3].

Educational Data Mining Methods

EDM not apply only data mining techniques Classification, clustering, and association analysis, but also apply methods and techniques drawn from the variety of areas related to EDM (statistics, machine learning, text mining, web log analysis, etc.). There are so many methods of educational data mining but all kind of methods lie in one of following categories

Prediction

The goal is to develop a model which can infer a single aspect of the data (predicted variable) from some combination of other aspects of the data (predictor variables). Types of predictions methods are classification, regression (when the predicted variable is a continuous value), or density estimation (when the predicted value is a probability density function).

Regression

Regression is an inherently statistical technique used regularly in data mining. Regression analysis establishes a relationship between a dependent or outcome variable and a set of predictors. Regression is supervised learning data mining technique. Supervised learning partitions the database into training and validation data. There are two type of regression technique.

Types

1. Linear regressions.
2. Non linear regressions.
3. Classification

Classification derives a model to determine the class of an object based on its attributes. A collection of records will be available, each record with a set of attributes. One of the attributes will be class attribute and the goal of classification task is assigning a class attribute to new set of records as accurately as possible. classifies a data item into some of several predefined categorical classes .The algorithm used for classification are:

- Decision tree
- Naive biased classification
- Generalized Linear Models (GLM)
- Super vector machine

Clustering

In clustering technique, the data set is divided in various groups, known as clusters. As per clustering phenomenon, the data point of one cluster and should be more similar to other data points of same cluster and more dissimilar to data points of another cluster. There are two ways of initiation of clustering algorithm: Firstly, start the clustering algorithm with no prior assumption and second is to start clustering algorithm with a prior postulate.

Outlier Detection

The goal of outlier detection is to discover data points that are significantly different than the rest of data. An outlier is a different observation (or measurement) that is usually larger or smaller than the other values in data. In EDM, outlier detection can be used to detect deviations in the learner's or educator's actions or behaviors, irregular learning processes, and for detecting students with learning difficulties. [2], [3] [7].

Phases of Education Data Mining

Educational Data Mining is concerned with translation of new hidden information from the raw data collected from educational systems. EDM generally consist of following phases:

The data is collected which is to be mined from different educational system resources i.e. from course management system (different institutes), E-learning environment, web based data (i.e. YouTube, twitter)which is relevant to students activities during learning process(i.e. their academic grades, students posts on social networking sites etc)

Educational Data Mining Process Phases

The first phase in the EDM process (not including pre-processing) is to discover relationships between data. It is accomplished by searching in a repository of data in an educational environment with the objective of finding consistent relationships between data variables. Various algorithms are used for finding such relationships some of them are classification, regression, clustering, factor analysis, social network analysis, association rule mining, and sequential pattern mining.

Validating Relationships

Discovered relationships must be validated in order to avoid over fitting.

Making Predictions

Relationships which are valid are used to make predictions about further events in the learning environment.

Decision Making

Predictions made in previous phase are used to support decision-making processes and in making policy decisions. During phases 3 and 4, data is often visualized or n some other way to make distillation of human judgment. Visualising Data, is one of the best practices in research [6] [10].

Goals for Educational Data Mining in Educational Field

1. **Predicting student's future learning behaviour:** With the use of student modelling, this goal can be achieved by creating student models that incorporate the learner's characteristics, including detailed information such as their knowledge, behaviors and motivation to learn.
2. **Discovering or improving domain models:** Through the various methods and applications of EDM, discovery of new and improvements to existing models is possible.
3. **Advancing scientific knowledge about learning and learners:** By building and incorporating student models, the field of EDM research and the technology and software used.
4. **Student Modeling:** User modeling in the educational domain incorporates such detailed information as student's characteristics or states such as knowledge, skills, motivation, satisfaction, meta-cognition, attitudes, experiences and learning progress, or certain types of problems that negatively impact their learning outcomes. The common objective here is to create or improve a student model from usage information.
5. **Predictive Modeling:** Predicting students' performance and learning outcomes. The objective is to predict a student's final grades or other types of learning outcomes (such as retention in a degree program or future ability to learn) based on data from course activities.
6. **Generating Recommendations:** The objective is to recommend students that content (or tasks or links) which is the most appropriate for them at the current time.
7. **Analyzing learner's behavior**: This takes on several forms: Applying educational data mining techniques to analyze learner behavior.
8. **Maintaining and improving courses**: The objective here is to determine how to improve courses (contents, activities, links, etc.), using information (in particular) about student usage and learning. Discovering or improving models that characterize the subject matter to be learned (e.g. math,

science, etc.), identify fruitful pedagogical sequences, and suggest how these sequences might be adapted to student's needs. Studying the effects of varied pedagogical enhancements on student learning.

9. **Learners:** To support a learner's reflections on the situation, to provide adaptive feedback or recommendations to learners, to respond to student's needs, to improve learning performance, etc.
10. **Educators:** To understand their student's learning processes and reflect on their own teaching methods, to improve teaching performance, to understand social, cognitive and behavioral aspects, etc.
11. **Administrators:** To evaluate the best way to organize institutional resources (human and material) and their educational system [2] [14].

Issues in Educational Data mining

A. **Lack of Data Interoperability:** Scalable Data management has become critical considering wide range of storage locations, data platform heterogeneity and a plethora of social networking sites. E.g: Metadata Schema Registry is a tool to enhance Meta data interoperability. So there is a need to design a model to classify/ cluster the data or find relationships. Examples of clustering applications are grouped students based on their learning and interaction patterns used in and grouping users for purposes of recommending actions and resources to similar users. It is possible to introduce Neuro-Fuzzy mining technique to remove the gap of data interoperability.

B. **Possibility of Uncertainty:** Due to the presence of uncertain errors, no model can predict hundred percent accurate results in terms of student modeling or overall academic planning.

C. **Educational data is incremental in nature**: Due to the exponential growth of data, the maintaining the data in data warehouse is difficult. To monitor the operational data sources, infer the student interest, intentions and its impact in a particular institution is the main issue. Another issue is the alignment and translation of the incremental educational data. It should focus on appropriating time, context and its sequence. Optimal utilization of computing and human resources is another issue of incremental.

D. **Research Expertise Relation between Student-Teacher:** In most of the higher Educational institutions (e.g. Engineering Institutions) final year students have a compulsory project work which are a research work based on their area of interest. Generally Supervisors are assigned as per availability and area of expertise in the respective department. But still it is not possible to assign all the students –supervisor with similar area of interest hence the result of the project is not applicable to real scenarios. There is need to find the relation between areas of interest, students' interest, applicability of the project/research and mining cross faculty interest [11] [13].

Applications of Education Data Mining

A. Predicting Student Performance - In student performance prediction, we predict the unknown value of a variable that defines the student. In educational sector, the mostly predicted values are student's performance, their marks, knowledge or score. Classification technique is used to combine individual items based upon quantitative traits or based upon training set of previously labeled items. Student's performance prediction is very popular application of DM in education sector. Different techniques and models are applied for prediction of student's performance like decision trees, neural networks, rule based systems, Bayesian networks etc. This analysis is helpful for someone in predicting student's performance i.e. prediction about student's success in a course and prediction about student's

final grade on the basis of features taken from logged data. Several regression techniques are used for prediction of student's marks such as linear regression to predict student's academic performance, stepwise linear regression to predict time spent by a student on a learning page and multiple linear regression for identification of variables that are helpful for predicting success in colleges courses and for prediction of exam results in distance education courses.

B. Enrolment Management - Enrolment management is frequently used in higher education to explain well-planned strategies and ways to shape the enrolment of college to meet planned goals. It is an organizational concept and also a systematic set of activities designed to allow educational institutions to exert more influence over student's enrolments. Such practices often include retention programs, marketing, financial aid awarding and admission policies.

C. Analysis and Visualization of Data - It is used to highlight meaningful information and support decision making. In the educational sector, for example, it can be helpful for course administrators and educators for analyzing the usage information and students‟ activities during course to get a brief idea of a student‟ s learning. Visualization information and statics are the two main methods that have been used for this task. Statistical analysis of educational data can give us information like where students enter and exit, the most important pages students browse, how many number of downloads of e-learning resources, how many number of different type of pages browsed and total amount of time for browsing of these different pages. It also provides information about reports on monthly and weekly user trends, usage summaries, how much material students will study and the series in which they study topics, patterns of studying activity, timing and sequencing of activities.

D. Grouping Students - In this case groups of students are created according to their customized features, personal characteristics, etc. These clusters/groups of students can be used by the instructor/developer to build a personalized learning system which can promote effective group learning. The DM techniques used in this task are classification and clustering. Different clustering algorithms that are used to group students are hierarchical agglomerative clustering, K-means and model-based clustering. A clustering algorithm is based on large generalized sequences which help to find groups of students with similar learning characteristics like hierarchical clustering algorithm which are used in intelligent e-learning systems to group students according to their individual learning style preferences discriminating features and external profiling features.

E. Predicting Students Profiling - Data mining can help management to identify the demographic, geographic and psychographic characteristics of students based on information provided by the students at the time of admission. Neural networking technique can be used to identify different types of students.

F. Planning and scheduling - Planning and scheduling is used to enhance the traditional educational process by planning future courses, course scheduling, planning resource allocation which helps in the admission and counseling processes, developing curriculum, etc. Different DM techniques used for this task are classification, categorization, estimation, and visualization. Decision trees, link analysis and decision forests have been used in course planning to analyze enrollee‟ s course preferences and course completion rates in extension education courses. Educational training courses have been planned through the use of cluster analysis, decision trees, and back-propagation neural networks in order to find the correlation between the course classifications of educational training. Decision trees and Bayesian models have been proposed to help management institutes to explore the probable effects of changes in recruitments, admissions and courses [4] [8]

Conclusion

Educational Data Mining (EDM) is an emerging multidisciplinary research area, in which methods and techniques for exploring data originating from various educational information systems have been developed..This paper focus about educational data mining, methods, goals and applications in educational data mining. The application of data mining methods in the educational sector is an interesting phenomenon. Data mining techniques in educational organizations help us to learn student performance, student behavior, designing course curriculum and to motivate students on various parameters.

References

1. J. Han, and M. Kamber, 2006. Data Mining Concepts and Techniques, Elsevier Publishers.
2. Ritu Gautam1 and Deepika Pahuja "A Review on Educational Data Mining ", IJSR International Journal of Science and Research, Vol-3, Issue-11, November-2014
3. Geeta Kashyap and Ekta Chauhan "Review on Educational data Mining techniques", ICRISEM-15
4. S. Lakshmi Prabha and Dr.A.R.Mohamed Shanavas, "Education Data Mining Applications", ORAJ Operation Research and Applications an International Journal,Vol.1, No.1,August 2014
5. Monika Goyal and Rajan Vohra, "Applications of Data Mining in Higher Education", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 2, No 1, March 2012.
6. Dina Abdulaziz Alhammadi, "Data Mining in Education- An Experimental Study", International Journal of Computer Applications Volume 62, No.15, January 2013.
7. Dr. Mohd Maqsood Ali, "Role of data mining in education sector", International Journal of Computer Science and Mobile Computing Vol. 2, Issue. 4, April 2013.
8. Dr. P. Nithya, B. Umamaheswari," A. Umadevi "A Survey on Educational Data Mining in Field of Education "International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 5 Issue 1, January 2016.
9. Mrs. M.S. Mythili, Dr. A.R.Mohamed Shanavas, "An Analysis of students" performance using classification algorithms", IOSR Journal of Computer Engineering, Volume 16, Issue 1, January 2014.
10. N. Upadhyay, V. Katiyar, A Survey on the Classification Techniques in Educational Data Mining, International Journal of Database Management System (IJDMS), 3(11), 2014, 725–728..
11. R.Jindal, M.D Borah, A Survey on Educational Data Mining and Research trends, *International Journal of Database Management System (IJDMS),* Vol.5, No.3, June 2013.
12. U. K. Pandey, and S. Pal, "A Data mining view on class room teaching language", (IJCSI) International Journal of Computer Science Issue, Vol. 8, Issue 2, pp. 277-282, ISSN:1694- 0814, 2011.
13. Abdulmohsen Algarni" Data Mining in Education" (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 6, 2016.
14. R. Baker et al., "Data mining for education," International encyclopedia of education, vol. 7, pp. 112–118, 2010.

PENETRATION TESTING METHODS, IMPORTANCE AND APPROACHES

V.Gayathri
Assistant Professor, Department of Networking,
Subbalakshmi Lakshmipathy College of Science, Madurai, Tamil Nadu, India

P.Ramya
Assistant Professor, Department of Networking,
Subbalakshmi Lakshmipathy College of Science, Madurai, Tamil Nadu, India

Abstract
Penetration testing is the methodology used for testing the computer system, network or Web application to predict vulnerabilities that an attacker could exploit in order to spoil the target. Penetration testing is also called pen testing which is mainly carried out by ethical hacker or white hat hacker. Pen tester acts like an attacker, analyses the system or network, predicts the vulnerability and safeguard the vulnerabilities in the system. Penetration testing will help to identify the main threats such as communication failure and e-commerce failure, loss of confidential information and authenticity, DNS, firewalls and passwords vulnerabilities. Penetration tester should develop objectives and goals, limitations and justification procedures based on the system or network weakness of the organization. This paper shows various pen testing methods for handling security issues in computer system and networking environment.
Keywords*: Penetration testing, DNS, Firewalls, passwords vulnerabilities.*

Introduction

A Network is a connection between two or more computer systems linked together which allows to transfer data between the computer systems. The connection between the nodes may be cable connection or wireless connection. In the network several devices are used for interconnection such as routers, hubs, switches, bridges and gateways etc. computer devices that originate, route and terminate the data are called network nodes. Nodes include hosts such as personal computers, phones, servers as well as networking hardware. Networks can be categorized based on topology, protocol and architecture. Topology is nothing but the geometric arrangement of a computer system. Common topologies are bus, star, and ring. The protocol defines a set of rules and signals use to communicate between the computers in the network. Most popular protocols for LANs is Ethernet. Different communication protocols are used for communications such as IEEE 802, wireless LAN, internet protocol suite, Asynchronous transfer mode, SONET. Networks can be broadly classified as either a peer-to-peer or client/server architecture. Networks are characterized by its physical capacity or its organizational purpose .Based on the geographical area networks can be categorized as wide-area networks (WAN),Personal area network(PAN), local-area networks (LAN)metropolitan-area networks (MAN)home-area networks (HAN), Enterprise private network, Virtual private network, Global area network . Virtual private network(VPN) networks are used for private organization communications ,connection between the nodes are carried by open connections or virtual circuits. VPNs are used to separate the traffic between different user communities in the underlying network with strong security features. VPN defined service level agreement (SLA) between the VPN customer and the VPN service provider. VPN has more complex topology than point-to-point network architecture. Intranets are the combination of networks with in the single organization .Internetwork is the interconnection of

multiple networks via a common routing technology. Internet is the worldwide communication network that uses the internet protocol suite to connect the devices over the world. Various applications of networks are World Wide Web, Video, Digital Audio. Recently emerging applications of networks are mobile computing, cloud computing, grid computing, wireless networks, telemedicine and security. Security plays a major role in the network environment. Networks vulnerabilities are mis-configuration of network devices, incomplete configuration of devices. Some of the attacks in the networks are man in the middle attack, rogue access point attack, session hijacking, Denial of service etc.

Why Penetration Testing?

Penetration testing plays a major role in estimating and maintaining security of a system or network. It helps you in finding out the loop holes by installing attacks. It consists of both script-based testing and human-based testing on networks. A penetration test not only reveals network security flaws and also provides risk assessment. Let's see what are the things can be done with the help of penetration testing:

- You can identify the threats plaster an organization's information assets.
- You can reduce an organization's security costs and also provide better Return On IT
- Security Investment (ROSI) helps for identifying and rectifying vulnerabilities and weaknesses.
- You can provide an organization with declaration: a thorough and comprehensive.
- Evaluation of organizational security covering policy, procedure, design and implementation.
- You can gain and sustain certification to an industry regulation (BS7799, HIPAA, etc.).
- You can agree best practices by compatible to legal and industry regulations.
- You can check and authenticate the efficiency of security protections and controls.
- It concentrations on high-severity vulnerabilities and highlights application–level security issues to development teams and management.
- It offers a comprehensive approach of preparation steps to prevent upcoming abuse.
- You can assess the effectiveness of network security devices such as firewalls, routers and web servers.
- You can practice it for changing or upgrading existing infrastructure of software, hardware, or network design.

Penetration Test Process

Let's start the actual process followed by test assistances or penetration testers. Identifying weaknesses present in system is the first most step in this process. Remedial action is taken on that vulnerability and same penetration tests are repeated again until system is non-vulnerable to all the tests [12]. We can categorize this process in following methods:

- Data collection
- Vulnerability assessment
- Actual exploit
- Result analysis and report preparation

1) Data collection: Various methods containing Google search are used to collect target system data. One can also take the web page source code analysis technique to collect more info about the system, software and plugin versions. There are many free tools and services which are needed for pen test is available in the market which provides you information like database or table names, DB

versions, software versions, hardware used and various third party plugins followed in the target system.

2) Vulnerability Assessment: Depends on the data collected in first step one can easily find out the security weakness in the target system. This helps penetration testers to do attacks using known entry points in the system.

3) Actual Exploit: This is critical step. It needs special skills and techniques to do attack on target system. Well knowledge penetration testers can use their skills to launch the attack on the system.

4) Result analysis and report preparation: After achievement of penetration tests detailed reports are arranged for taking corrective actions. All identified vulnerabilities and suggested corrective methods are listed in these reports. You can able to change the vulnerability report format (HTML, XML, MSWord or PDF) as per the organization needs.

Penetration Testing Phases

1. Pre attack phase
2. Attack phase
3. Post attack phase

Pre Attack Phase

This phase is encouraged on collecting more information about the target organization or network to be criticized. This can be done by either non-invasive or invasive.

Attack Phase

The information gathered in the pre-attack phase produces the basis of the attack plan. Before deciding the attack policy, the tester may choose to carry out an offensive information gathering process like scanning is done.

Post Attack Phase

This is a critical part of the testing process, as the tester wants to restore the network to its original state. This undergoes clean-up of testing processes and exclusion of vulnerabilities created (not before existed originally), exploits created etc.

Testing Point

Every penetration test have a start and end-point, unrelated of whether it is zero knowledge or partial knowledge test. How does a pen test team or an organization regulate this? Though providing a penetration - testing team with information such as the exact configuration of the firewall used by the target network may faster the testing; it can work harmfully by providing the testers with an unworkable advantage. [12]Goal of the penetration effort is to find as much vulnerability as possible, it strength for white box testing and share as much information as possible with the testers.. Therefore, stability must be reached between support the testing team in accompanying their test faster and providing a more realistic testing environment by limiting information. Some organizations may select to get the initial pen test examined by a second pen test team so that there is a third party declaration on the results obtained.

Types of Penetration Testing

The following tests originates fall under penetration testing:

Gray - Box Penetration Testing

[13] In gray-box penetration testing, the test is accompanied with limited knowledge about infrastructure, defense mechanism, and communication channels of the target. It is model of those

attacks that is performed by the insider or outsider with limited right of entry privileges. In this case, administrations would prefer to provide the pen testers with partial familiarity or information that hackers could find such as domain name server. It helps to save time and expenses of the organization. In gray-box testing, pen testers interact with system and network administrators.

White - Box Penetration Testing

[13] In white-box penetration testing, the test is lead with full knowledge of infrastructure, defense mechanism, and communication channels of the target on which test is to be conducted. This test pretends the insider attacker who has full rights and unrestricted access to the target system. This type of penetration test is conducted when the organization wants to evaluate its security against a specific kind of attack or a specific objective. In this case, the complete information is given to the pen testers about the target. The information given may include network topology documents, asset inventory, and valuation information. Typically, an organization would suitable for this when it wants a complete examination of its security.

Announced / Unannounced Testing

[5] Announced testing is an effort to access and recover pre-identified flag file(s) or to cooperative systems on the client network with the full support and knowledge of the IT staff. Such testing audits the existing security infrastructure and individual systems for the existence of vulnerabilities. Creating a team-oriented location in which members of the organization's security staff are part of the penetration team allows for a embattled attack against the most valuable hosts. Unannounced testing is an effort to access and retrieve pre-identified flag file(s) or to compromise systems on the client network with the alertness of

Internal Network Penetration Testing

In internal network penetration testing, all the possible internal network flaws are identified and simulated as if a real attack has taken place. Various methods used for the internal network penetration testing are:

- Internal Network Scanning
- Port Scanning
- System Fingerprinting

Router and Switches Penetration Testing

[15] Router switches penetration is carried out to determine:

- End to end router security
- Bandwidth and speed of the internet connection

Firewall Penetration Testing

[10] Firewall penetration testing is one of the most useful methods in analyzing security effectiveness. Through this method, you can identify how secure your firewall network is against the attacks performed by network intruders.

IDS Penetration Testing

[3] An intrusion detection system (IDS) can be software or hardware. IDS penetration testing helps you to test the strength of the IDS. It can be performed with the help of tools such as IDS informer, an evasion gateway, etc.

Wireless Network Penetration Testing

Wireless networks are more economical than wired networks. Though wireless networks are cheaper, there are various risks associated with them. A wireless network is less protected than a wired one. Therefore, wireless networks must be tested strictly and the respective security enhancements must be applied.

Denial-of-Service penetration Testing

[4] The main purpose of a denial-of-service (DOS) attack is to slow down the website or even to crash it by sending too many requests, more than a particular server can handle. If the attacker knows the details of the server and its technical specifications, it becomes more vulnerable. Sometimes DOS is done on a trial and error basis. So the penetration tester must check how much the website or server can withstand. It is also necessary to provide an alternative way to react to the situation when the limit exceeds.

Password Cracking Penetration Testing

[20] Passwords are used to protect computer resources from illegal access. Password cracking penetration testing identifies the vulnerabilities associated with password management. This helps in avoiding various kinds of malicious attacks such as brute force attacks, hybrid attacks, and dictionary attacks, etc.

Social Engineering Penetration Testing

[14] Social engineering is a method used by attackers to get vital information of a company. Attackers especially target individuals within the organization to gather as much information as possible about the company. This is completely documented and then the employees are educated about possible social engineering and cautioned about various threats.

SQL Injection Penetration Testing

[19] The penetration tester should perform SQL injection penetration testing on the application in order to find out vulnerabilities in the application. The pen tester should try to simulate different types of SQL injection attacks to find out the possible vulnerabilities.

Physical Security Penetration Testing

Here the penetration testing tries to gain physical access to the organizational resources before, during, and after the business hours. All the physical security controls must be properly tested.

Database Penetration Testing

[17] In this process, a penetration tester tries to directly access data contained in the database or indirectly accessing the data through triggers or stored procedures executed by a database engine. This method helps in avoiding unauthorized access of data.

VPN Penetration Testing

[16] Sometimes, workers are allowed to work from home or remotely and in such situations, there is lot of security issues associated with VPN. So the penetration attempts to gain access to the VPN through an isolated endpoint or VPN tunnel and check the vulnerabilities.

Cloud Penetration Testing

Cloud computing systems are widespread today. There are risks associated with cloud computing. The organizations must figure out these risks and apply proper security mechanisms to protect against potential risks. To find out the vulnerabilities in a cloud-based application, conduct a penetration test on the cloud.

Virtual Machine Penetration Testing

An assailant can exploit the virtual machine security fault by running malevolent code on the virtual machine. The pen tester needs to find out the vulnerabilities in the VM by simulating the actions of an attacker before a real attack occurs.

Virus and Trojan Detection

Virus and Trojans are the most prevalent malevolent software today. Once on the system and networks, these are very dodgy. Early detection of viruses and Trojans is very important engineering or extra means.

Data Leakage penetration Testing

Penetration testing of data leakage aids in the following means

- Avoiding confidential information to going out in the place of market or to competitors
- Permits growing internal compliance level for data protection

Conclusion

Penetration testing is more helpful to find the security vulnerabilities and weakness in the system and networks. Based on the testing results we can able to safeguard the systems and rectify the weakness and holes in the system. It helps to secure the systems from attack. Pen tester plays a role like attacker and finds the weakness and prepares the security measures for that in order to prevent the attack of the unauthorized person (attacker).This paper gives the efficient knowledge about pen testing and importance of pen testing in any secured organization.

Table 1: Comparison of Various Pen testing Methods

Parameter	Data Traffic	Integrity	Data Security	Authentication	Authorization	Scalability	Data Availability	Knowledge of Pen tester
Black box penetration testing	×	√	×	×	×	×	×	▲
Gray - box penetration testing	√	√	×	√	√	×	×	▲
White - box penetration testing	√	√	√	√	√	×	√	●
Announced / unannounced testing	√	√	√	×	×	×	√	▲
Internal Network Penetration Testing	√	√	√	√	×	√	×	●
Router and Switches Penetration Testing	√	√	√	×	×	√	√	●
Firewall Penetration Testing	√	√	√	√	√	√	×	●

IDS Penetration Testing	√	√	√	√	√	√		●
Wireless Network Penetration Testing	√	×	√	√	×	√	√	●
Denial-of-Service penetration Testing	√	√	√	√	√	√	√	●
Password Cracking Penetration Testing	×	√		√	√	×	×	▲
Social Engineering Penetration Testing	×	√	√	√	√	×	×	▲
SQL Injection Penetration Testing	√	√	√	√	√	×	√	●
Virtual Machine Penetration Testing	√	√	√	√	√	√	√	●
Physical Security Penetration Testing	×	√	√	√	√	√	×	●
Database Penetration Testing	√	√	√	√	√	√	√	●
Virus and Trojan Detection	√	√	√	√	√	×	×	●
Data Leakage Penetration Testing	√	√	√	√	√	×	√	●
VPN Penetration testing	√	√	√	√	√	√	√	●
Cloud Penetration Testing	√	√	√	√	√	√	√	●

√ - Issue is addressed; ● - High; ▲ - Low; × - Issue not addressed

References

1. Tao Lu, Jie Chen," Research of Penetration Testing Technology in Docker Environment " in Advances in Engineering Research, volume 141, 2017.
2. Paschal A. Ochang, Philip Irving, "Security Analysis of VoIP Networks Through Penetration Testing" in International Conference on Information and Software Technologies 2017.
3. Nicholas J. Puketza," IDS PENETRATION TESTING", Department of Computer Science University of California-2013.
4. Andreas Falkenberg, "A New Approach towards DoS Penetration Testing on Web Services",2013 IEEE 20th International Conference on Web Services
5. Daniel Dalalana Bertoglio, "Overview and open issues on penetration test" ,in Journal of the Brazilian Computer Society 2017.
6. "Penetration Testing: Assessing Your Overall Security Before Attackers Do". SANS Institute. Retrieved 16 January 2014.
7. Kevin M. Henry. Penetration Testing: Protecting Networks and Systems. IT Governance Ltd. ISBN 978-1-849-28371
8. Penetration testing is the simulation of an attack on a system, network, piece of equipment or other facility, with the objective of proving how vulnerable that system or "target" would be to a real attack.
9. "Penetration Test Guidance Special Interest Group PCI Security Standards Council", March 2015.
10. Reto E. Haeni, r.haeni@cpi.seas.gwu.edu, The George Washington University, Cyberspace Policy Institute, 2033 K Str. Suite 340 N, Washington DC 20006, Washington DC, January 1997
11. Liwen He Nikolai Bode, "Network Penetration Testing" in springer.
12. G. Bacudio "An Overview of Penetration Testing", in International Journal of Network Security & Its Applications (IJNSA), Vol.3, No.6, November 2011.
13. "Study Paper on Penetration Testing Methodologies" Telecommunication Information Center -2014

14. "Two methodologies for physical penetration testing using social engineering" in ACSAC '10 Proceedings of the 26th Annual Computer Security Applications Conference.
15. Philip R. Moyer E. EugeneSchultz ,"A systematic methodology for firewall " penetration testing in Network Security Volume 1996, Issue 3, March 1996, Philip R.MoyerE. EugeneSchultz.
16. Ron Gula, "BROADENING THE SCOPE OF PENETRATION-TESTING TECHNIQUES" in Intrusion Detection Products Enterasys Networks.
17. Andrey Petukhov, Dmitry Kozlov, "Detecting Security Vulnerabilities in Web Applications Using Dynamic Analysis with Penetration Testing " in Application security conference
18. C.C. Chen, J. Ruuskanen, W. Pilacinski & K. Willeke," FILTER AND LEAK PENETRATION CHARACTERISTICS OF A DUST AND MIST FILTERING FACEPIECE" in American Industrial Hygiene Association Journal Volume 51.
19. Nuno Antunes , Marco Vieira "Comparing the Effectiveness of Penetration Testing and Static Code Analysis on the Detection of SQL Injection Vulnerabilities in Web Services" in Dependable Computing, 2009. PRDC '09. 15th IEEE Pacific Rim International Symposium.

A SURVEY ON FILTERS USED IN SINGLE IMAGE DEHAZING

M.Ramesh Kanthan
Research Scholar, Bharathiar University, Coimbatore, Tamil Nadu, India

Dr.S.Naganandini Sujatha
Assistant Professor, Government Arts College for Women, Nilakottai, Tamil Nadu, India

Abstract
The problem focused in the area of image processing is the restoration of the images those are corrupted due to various weather degradations. Images of outdoor scenes captured in a bad weather conditions contain atmospheric degradation such as haze, fog, smoke. The dark channel prior is a highly used method using a kind of statistics of outdoor haze-free images. It is based on a key observation that most local patches in outdoor haze-free images contain some pixels whose intensity is very low in at least one color channel. Using this prior with the haze imaging model, directly estimate the thickness of the haze by transmission map and recover a high-quality haze-free image. This paper proposes the concept of using different filtering techniques to refine the transmission map in order to get quick and accurate results. In this paper, we elaborate single image dehazing by combining different filters. By analyzing the tradeoffs of this approach, we propose an effective scheme to adapt the parameters. Experiments and comparisons show that our survey generates satisfactory dehazed results with low computation.
Key Words*: Dehazing, dark channel prior, bilateral filter, Guided filter, transmission map.*

Introduction

Dehazing is the process to remove haze effects in captured images and reconstruct the original colors of natural scenes, which will be a useful pre-processing for input images and required for receiving high performance of the vision algorithm. However, haze removal is a challenging problem since the degradation is spatial-variant, it depends on the unknown scene depth. There are several method proposed for fog removal based on contrast enhancement .Few method analyze and process the image based solely on the information from the image or multiple observation of the same scene .Commonly used global contrast enhancement techniques are Linear mapping, Histogram stretching, Histogram equalization and Gamma correction. Other contrast enhancement techniques that amplify local variations in Intensity as per pixel fashion in order to increase the accuracy and sharpening of the haze free process. Laplacian pyramid, Wavelet decomposition, single scale unsharp mask filter, multi scale bilateral filter and Guided filter can be found in the local processing techniques.

This proposed approach is based on the fusion strategy and it has been derived from the original hazy image inputs by applying a different filtering approach. It combines several intermediate results which containing most significant features, obtain from single foggy image, based on the local processing techniques. The fusion enhancement technique estimates perceptual based qualities known as the weight maps for each pixel in the image. Single image dehazing often suffers from the problem of ambiguity between image color and depth. That is, a clean pixel may have the same color with a fog-contaminated pixel due to the effects of hazes. The proposed algorithm focuses on that limitation also.

Literature Survey

Under bad weather conditions, however, the contrast and color of images are drastically altered or degraded. Hence, it is imperative to remove weather effects from images in order to make vision systems more reliable. Enhancement of foggy image is a challenge due to the complexity in recovering luminance and chrominance while maintaining the color fidelity. During enhancement of foggy images,

it should be kept in mind that over enhancement leads to saturation of pixel value. Thus, enhancement should be bounded by some constraints to avoid saturation of image and preserve appropriate color fidelity. Hence basic challenge is to nullify the whitening effect thereby improving the contrast of the degraded image. Manoj Alwani and Anil Kumar Tiwaria [1] present a contrast enhancement algorithm for degraded color image. S. G. Narasimhan and S. K. Nayar, [2] proposed a physics-based model that describes the appearances of scenes in bad weather conditions. J. P. Oakley and H. Bu [4] proposed a method for determination of airlight level in digital images. The method involves the minimization of a scalar global cost function and no region segmentation is required.

Proposed System

Atmosphere contains the fog and haze particles so that the color and contrast of the images are drastically degraded. The degradation level increases with the distance from the camera to the object. The removal of haze from the captured hazy images needs to estimate the depth of the haze. The initial works for the haze removal uses multiple input images those have been derived from the filtered output.

The general prototype of the image dehazing techniques is shown in figure 1.

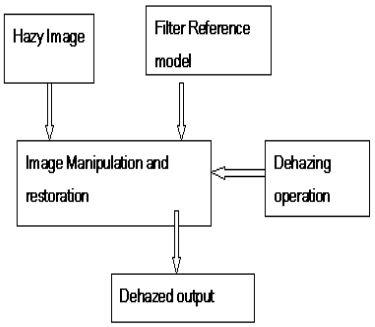


Figure 1: Image Dehazing Technique Prototype

Bilateral Filter

The bilateral filter computes the filter output at a pixel as a weighted average of neighboring pixels. It smoothes the image while preserving edges. Due to this nice property, it has been widely used in noise reduction, HDR compression, multi-scale detail decomposition, and image abstraction. It is generalized to the joint bilateral filter, in which the weights are computed from another guidance image rather than the filter input. The joint bilateral filter is particular favored when the filter input is not reliable to provide edge information, e.g., when it is very noisy or is an intermediate result. The joint bilateral filter is applicable in flash/no-flash denoising, image upsampling, and image deconvolution.

Optimization-based Image Filtering

A series of approaches optimize a quadratic cost function and solve a linear system, which is equivalent to implicitly filtering an image by an inverse matrix. In Image segmentation and colorization the affinities of this matrix are Gaussian functions of the color similarities. In image matting, a matting Laplacian matrix is designed to enforce the alpha matte as a local linear transform of the image colors. This matrix is also applicable in haze removal .The weighted least squares (WLS) filter adjusts the matrix affinities according to the image gradients and produces a halo-free decomposition of the input image. Although these optimization-based approaches often generate high quality results, solving the corresponding linear system is time-consuming. It has been found that the optimization-based filters are closely related to the explicit filters.

Guided Filter

A general linear translation-variant filtering process, which involves a guidance image I, an input image p, and an output image q. Both I and p are given beforehand according to the application, and they can be identical. The filtering output at a pixel i is expressed as a weighted average

$$q_i = \sum_j W_{ij}(I) p_j, \tag{1}$$

where i and j are pixel indexes. The filter kernel Wij is a function of the guidance image I and independent of p. This filter is linear with respect to p.

Edge-Preserving Filtering

An example of the guided filter with various sets of parameters is used in edge preserving operation. We can see that it has the edge-preserving smoothing property. This can be explained intuitively as following. Consider the case that I = p. It is clear that if _ = 0, then the solution to (4) is ak = 1 and bk = 0. If _ > 0, we can consider two cases:

Case 1: "Flat patch". If the image I is constant in !k, then it is solved by

ak = 0 and bk = ¯pk;

Case 2: "High variance". If the image I changes a lot within !k, then ak becomes close to 1 while bk is close to 0. We have that if a pixel is in the middle of a "high variance" area, then its value is unchanged, whereas if it is in the middle of a "flat patch" area, its value becomes the average of the pixels nearby.

Gradient Preserving Filtering

Though the guided filter is an edge-preserving smoothing filter like the bilateral filter, it avoids the gradient reversal artifacts that may appear in detail enhancement and HDR compression. Given the input signal (black), its edge-preserving smoothed output is used as a base layer (red). The difference between the input signal and the base layer is the detail layer (blue). It is magnified to boost the details. The enhanced signal (green) is the combination of the boosted detail layer and the base layer. The shape of the edge is well maintained in the recombined layer.

Relation to the Matting Laplacian Matrix

The guided filter can not only be used as a smoothing operator. It is also closely related to the matting Laplacian matrix. This casts new insights into the guided filter and inspires some new applications. In a closed-form solution to matting .The matting Laplacian matrix is derived from a local linear model. Unlike the guided filter which computes the local optimal for each window, the closed-form solution seeks a global optimal.

Visibility Enhancement

Estimating the Atmospheric Light

In most of the previous single image methods, the atmospheric light **A** is estimated from the most haze-opaque pixel. For example, the pixel with highest intensity is used as the atmospheric light and is furthered refined. Among these pixels, the pixels with highest intensity in the input image **I** is selected as the atmospheric light.

Estimating the Transmission

Here, we first assume that the atmospheric light **A** is given. We will present an automatic method to estimate the atmospheric light. We further assume that the transmission in a local patch $\Omega(\mathbf{x})$ is constant and the patch's transmission as $\tilde{t}(\mathbf{x})$. Taking the min operation in the local patch on the haze imaging Equation

$$\tilde{t}(\mathbf{x}) = 1 - \min_{c}(\min_{\mathbf{y}\in\Omega(\mathbf{x})}(\frac{I^c(\mathbf{y})}{A^c})). \quad (2)$$

Recovering the Scene Radiance

With the transmission map, we can recover the scene radiance according to Equation (3). But the direct attenuation term **J**(**x**)t(**x**) can be very close to zero when the transmission t(**x**) is close to zero. The directly recovered scene radiance **J** is prone to noise. Therefore, we restrict the transmission t(**x**) to a lower bound t0, which means that a small

$$\mathbf{I}(\mathbf{x}) = \mathbf{J}(\mathbf{x})t(\mathbf{x}) + \mathbf{A}(1 - t(\mathbf{x})), \quad (3)$$

Experimental Results

Compared to the existing work the main advantage of combining dark channel prior and guided filter to dehaze the images lies in its low computational cost. On a laptop with a 2.2 GHz Intel Core 2 Duo CPU, our C++ implementation takes about 4 seconds to process a 1-mega pixel image. By virtue of its exact O(N) time complexity, the running time of our algorithm becomes tolerable for many applications. Some dehazed results of our work and comparisons with other existing algorithm are shown in Fig. 2. Since the recovered haze-free images usually look quite dim, we enhance their brightness for better display. As can be seen, our approach is capable of unveiling the details for the inputs, which gives results that comparable to He et al.'s work. However, guided image filtering is actually approximation of soft matting, as proven refining the transmission map with guided filter.

Figure 2: (i) Input (ii) existing method (iii) output

Conclusion

In this paper we have described that the haze layer present in the captured input mage is dependent on the scene depth and it is variant in nature. Also in this paper we have addressed different method in which the haze can be estimated from the captured hazy images and after estimating the depth map and using the image formation model a better and improved haze free image can be recovered. The main benefit of using guided filter to refine the transmission lies in its low computational cost; it also generates comparable dehazed results with existing work.

References

1. Jyoti Sahu "Design a new Methedology for Removing Fog from Image", IJACR Journal, Volume-2 Number-4Issue-7 Dec-2012.
2. A. Levin, D. Lischinski, and Y. Weiss. A closed form solution to natural image matting. CVPR, 1:61–68, 2006
3. S. G. Narasimhan and S. K. Nayar. Contrast restoration of weather degraded images. PAMI, 25:713–724,2003.

4. Dongjun Kim, and Hanseok Ko, "Enhancement of Image Degraded by Fog Using Cost Function Based on Human Visual Model", Proceedings of IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems Seoul, Korea, August 20 - 22, 2008.
5. Manoj alwani and Anil kumar tiwaria, "contrast enhancement based algorithm to improve visibility of colored foggy images" Recent advances in business administration.
6. S. G. Narasimhan and S. K. Nayar, "Contrast restoration of weather degraded images, "IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, pp. 713-724, 2003.
7. K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," in Proc. IEEE Conf. on Comp. Vis. Pat. Rec., June 2009.
8. K. He, J. Sun, and X. Tang, "Guided image filtering," in Proc. Europ.Conf. on Comp. Vis., Sep. 2010.
9. M. van Herk, "A fast algorithm for local minimum and maximum filters on rectangular and octagonal kernels," Pat. Rec. Lett., vol. 13, pp. 517- 521, July 1992.
10. R. Fattal, "Single image dehazing," ACM SIGGRAPH '08, 2008.
11. R. Tan, "Visibility in bad weather from a single image," in Proc. IEEE Conf. on Comp. Vis. Pat. Rec., June 2008.
12. S. G. Narasimhan and S. K. Nayar, "Vision and the atmosphere," Int'lJ. of Comp. Vis., vol. 48, pp. 233-254, 2002.

MINING SEQUENCE DATA AND TIME SERIES DATA

C.Sulochana, M.Sc., M.Phil.,
Assistant Professor in CS Department
Madurai Sivakasi Nadars Pioneer Meenakshi Women's College, Poovanthi

Abstract
Sequential and time series data mining remains an essential problem. Despite progress in other related fields, how to efficiently cluster, classify and predict the trends of these data is still an important topic. A particularly challenging problem is the noise in time series data. It is an open issue to tackle. Many time series used for predictions are contaminated by noise, making it difficult to do accurate short-term and long-term predictions. Examples of these applications include the prognosis of financial time series and seismic time series. Although signal processing techniques, such as wavelet analysis and filtering, can be applied to remove the noise, they often introduce straggle in the filtered data. Such lags reduce the accuracy of predictions because the predictor must overcome the delay before it can predict into the future. Existing data mining methods also have difficulty in handling noisy data and learning meaningful information from the data.
Keywords: *cluster, wavelet analysis, filtering, sequential and time series data mining.*

Introduction

The characteristics of the stream, time-series, and sequence data are unique, that is, large and endless. It is too large to get an exact result; this means an almost correct output will be achieved. The classic data-mining algorithm should be extending, or a new algorithm needs to be designed for this type of the dataset. A time-series database composed of sequences of values or occurrence obtained over repeated measurements of time. A sequence database consists of set of ordered elements or events, recorded with or without a concrete notion of time. Sequential pattern mining is the discovery of frequently occurring ordered events or subsequences as recurring design.

Proposed System

Wavelet is best for nonstationary signal analysis, here original signal attach with mother wavelet signal and then makes inspection, means Mother wavelet signal make Zoom of signal, then study, so the result is coming best in comparison to other techniques. But they often introduce lags in the filtered data. Such delay reduces the accuracy of prognosis because the predictor must overcome the lags before it can predict into the future.

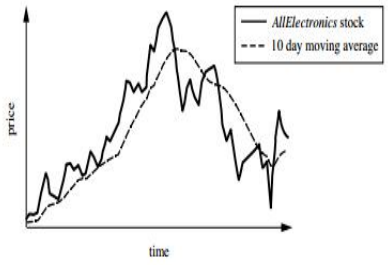
Mining Time Series Data

A time-series database consists of sequences of events obtained over repeated measurements of time. The values are typically measure at equal time intervals (e.g., hourly, daily, weekly). Time-series databases are rich in many applications, such as stock market analysis, economic and sales forecasting, budgetary analysis, utility studies, inventory studies, yield projections, workload projections, process and quality control, observation of natural phenomena (such as atmosphere, temperature, wind, earthquake), scientific and engineering experiments, and also for medical treatments. A time-series database is also a sequence database. However, a sequence database is any database that consists of sequences of ordered events, with or without concrete notions of time.

Trend Analysis

A time series involving a variable Y, representing, say, the daily closing price of a share in a stock market, can be viewed as a function of time t, that is, Y = F(t).Such a function can be illustrated as a time-series graph, as shown in Figure 1, which describes a point moving with the passage of time.

In general, there are two goals in time-series analysis: (1) modeling time series (i.e., to gain insight into the mechanisms or underlying forces that generate the time series), and (2) forecasting time series (i.e., to predict the future values of the time-series variables). Trend analysis consists of the following four major components or movements for characterizing time-series data:



Time-series data of the stock price of *AllElectronics* over time. The *trend* is shown with a dashed curve, calculated by a moving average.

Figure 1: Time Series Graph

- **Trend or long-term movements:** These indicate the general direction in which the time-series graph is moving over a long interval of time. These motions are displayed by a trend curve, or a trend line. For example, the trend curve of Figure 1 is indicated by a dashed curve. Typical methods for determining a trend curve or trend line include the weighted moving average method and the least squares method.
- **Cyclic movements or cyclic variations**: These refer to the cycles, that is, the long-term oscillations about a trend line or curve, which may or may not be periodic. That is, the series need not necessarily follow similar patterns after equal intervals of time.
- **Seasonal movements or seasonal variations**: These are systematic or calendar related. Samples include events that recur annually, such as the sudden increase in sales of chocolates and flowers of department store items before Christmas. The observed increase in water consumption in summer due to warm weather is another example. In these examples, seasonal movements are the identical or nearly identical patterns that a time series appears to follow during corresponding months of successive years.
- **Irregular or random movements:** These characterize the sporadic motion of time series due to unknown or chance events, such as labor disputes, floods, or announced personnel changes within companies.

Data Reduction and Transformation Techniques

Due to the tremendous size and high-dimensionality of time-series data, data reduction often serves as the first step in the time-series analysis. Data reduction leads to not only much smaller storage space but also much faster processing. Strategies for data reduction include attribute subset selection (which removes irrelevant or redundant attributes or dimensions), dimensionality reduction (which typically employs signal processing techniques to obtain a reduced version of the original data), and numerosity trimming (where data are replaced or estimated by alternative, smaller representations, such as histograms, clustering, and sampling). Because time series can be view as data of very high dimensionality where each point of time can be viewed as a dimension, dimensionality chopping is our main concern here. For example, to compute correlations between two time-series curves, the reduction of the time series from the length (i.e., dimension) n to k may lead to a reduction from O(n)to O(k)in computational complexity. If k << n, the complexity of the computation will be reduced.

Query Languages for Time Sequences

A time-sequence query language should be able to specify not only simple similarity queries like "Find all of the sequences similar to a given subsequence Q," but also sophisticated queries like "Find all of the sequences that are similar to some sequence in class C1, but not similar to any sequence in class C2." Moreover, it should be able to support various kinds of queries, such as range queries and nearest-neighbor queries. A category of time-sequence query language is a shape definition language. It allows users to define and query the overall shape of time sequences using human-readable series of sequence transitions or macros, while ignoring the specific details.

Mining Sequence Patterns in Transactional Databases

A sequence database consists of sequences of ordered elements or events, recorded with or without a concrete notion of time. There are many applications involving sequence data. Typical examples include customer shopping sequences, Web click streams, biological sequences, sequences of events in science and engineering, and in natural and social developments. In this section, we study sequential pattern mining in transactional databases.

Sequential Pattern Mining: Concepts and Primitives

Sequential pattern mining is the mining of frequently occurring ordered events or subsequences. An example of a sequential pattern is "Customers who buy a Canon digital camera are likely to buy an HP color printer within a month." For retail data, sequential designs are useful for shelf placement and promotions. This industry, as well as telecommunications and other businesses, may also use sequential motifs targeted marketing, customer retention, and many other tasks. Other areas in which these can be applied include Web access pattern analysis, weather prediction, production processes, and network intrusion detection. Notice that most studies of sequential pattern mining concentrate on categorical(or symbolic) designs.

Let's establish some vocabulary for our discussion of sequential pattern mining. Let $I = \{I_1, I_2,..., I_p\}$ be the set of all items. An item set is a nonempty set of items. A sequence is an ordered list of events. A sequence s is denoted the $e_1,e_2,e_3 \cdots ei$, where event e_1 occurs before e_2, which occurs before e_3, and so on. Event e_j is also called an element of s. In the case of customer purchase data, an event refers to a shopping trip in which a customer bought items at a certain store. The event is thus an item set, that is, an unordered list of items that the customer purchased during the trip. The item set (or event) is denoted $(x_1x_2 \cdots x_q)$, where x_k is an item. For brevity, the brackets are omitting if an element has only one item, that is, element (x) is written as x. Suppose that a customer made several shopping trips to the store. These ordered events form a sequence for the customer. That is, the customer first bought the items in s1, then later bought. The items in s2, and so on. An item can occur at most once in the event of a sequence, but can occur multiple times in different events of a sequence. The number of instances of items in a sequence is called the length of the sequence. A sequence with length l is called an l-sequence. A sequence $\alpha = ha_1a_2 \cdots a_ni$ is called a subsequence of another sequence $\beta = hb_1b_2 \cdots b_mi$, and β is a supersequence of α, denoted as α v β, if there exist integers $1 \le j1 < j2 < \cdots < jn \le m$ such that $a_1 \subseteq b_j1$, $a_2 \subseteq b_j2$, . . . , $a_n \subseteq b_jn$. For example, if α = h(ab), di and β = h(abc), (de)i, where a, b, c, d, and e are items, then α is a subsequence of β and β is a supersequence of α.

A sequence database, S, is a set of tuples, SID, si, where SID is a sequence ID and s is a sequence. For our example, S contains sequences for all customers of the store. A tuple SID, si is said to contain a sequence α, if α is a subsequence of s. The support of a sequence α in a sequence database S is the

number of tuples in the database containing α, that is, supportS (α) = | {SID, si|(SID, si ∈ S)∧(α v s)} |. It can be denoted as support(α) if the sequence database is clear from the context. Given a positive integer min sup as the minimum support threshold, a sequence α is frequent in sequence database S if supportS (α) ≥ min sup. That is, for sequence α to be frequent, it must occur at least min sup times in S. A persistant sequence is called a sequential pattern. A sequential pattern with length l is called an l-pattern.

GSP: A Sequential Pattern Mining Algorithm Based On Candidate Generate-And-Test

GSP (Generalized Sequential Patterns) is a sequential pattern mining method that was developed by Srikant and Agrawal in 1996. It is an extension of their formative algorithm for recurrent item set mining, known as Apriori. GSP uses the downward-closure property of sequential patterns and adopts a multiple-pass, candidate generate-and-test approach. The algorithm is outlining as follows. In the first scan of the database, it finds all of the repeated items, that is, those with minimum support. Each such item yields a 1-event sequence consisting of that item. Each subsequent pass starts with a seed set of sequential designs—the set of sequential motifs found in the previous pass. This seed set is used to generate new potentially frequent designs, called candidate sequences. Each candidate sequence contains one more item than the seed sequential pattern from which it was generated (where each event in the pattern may contain one or multiple items). Recall that the number of instances of items in a sequence is the length of the sequence. So, all of the candidate sequences in a given pass will have the same length. We refer to a sequence with length k as a k-sequence. Let Ck denote the set of candidate k-sequences. A pass over the database finds the support for each candidate k-sequence. The candidates in Ck with at least min sup form Lk, the set of all frequent k-sequences. This set then becomes the seed set for the next pass, k+1. The algorithm terminates when no new sequential pattern is found in a pass, or no candidate sequence can be generated.

Conclusion

Since its conception in the late 1980s, data mining has achieved tremendous success. Many new problems have emerged and have been solved by data mining researchers. However, there is still a lack of timely exchange of topics in the community as a whole. So we simply solve one problem according to a reference of Mining time series and sequence data books.

References

1. Data Mining in Time Series Databases
2. Mark Last (Ben-Gurion University of the Negev, Israel), Abraham Kandel (Tel-Aviv University Israel & University of South Florida, USA), Horst Bunke (University of Bern, Switzerland) data-mining.philippe-fournier-viger.com
3. A introduction to sequential pattern matching by Phillippe-Fournier.
4. Mining Time series data by Jessica
5. R. Agrawal and R. Srikant, “Mining Sequential Patterns”. In Proceedings of the 11th International Conference on Data Engineering, pp. 3-14, Taipei, Taiwan, 1995
6. M. Garofalakis, R. Rastogi, and K. Shim, "SPIRIT: Sequential pattern mining with regular expression constraints," VLDB'99, 1999
7. J. Han and M. Kamber, “Data Mining: Concepts and Techniques,” Morgan Kaufman Publishers, 2001.

8. Srikant R. and Agrawal R., —Mining sequential patterns: Generalizations and performance improvements, Proceedings of the 5th International Conference Extending Database Technology, 1996, 1057, 3-17.
9. J.Pei, J.Han, B.MortazaviAsl, J.Wang, H.Pinto, Q.Chen, U.Dayal and M.-C.Hsu, "Mining sequential patterns by pattern-growth: The Prefix Span approach," IEEE Transactions on Knowledge and Data Engineering, vol.16, no.11, 2004, pp. 1424-1440.
10. Nizar R. Mabroukeh and C. I. Ezeife, "A Taxonomy of Sequential Pattern Mining Algorithms," ACM Computing Surveys, Vol. 43, No. 1, Article 3, Publication date: November 2010.
11. M. Zaki, "SPADE: An efficient algorithm for mining frequent sequences," Machine Learning, 2001.
12. S. Parthasarathy, M. Zaki, M. Ogihara, and S. Dwarkadas, "Incremental and interactive sequence mining," In Proc. of the 8th Int.Conf. On Information and Knowledge Management (CIKM'99), Nov1999.
13. Han J., Dong G., Mortazavi-Asl B., Chen Q., Dayal U., Hsu M.-C., Freespan: Frequent pattern-projected sequential pattern mining, Proceedings 2000 Int. Conf. Knowledge Discovery and Data Mining (KDD'00), 2000, pp. 355-359.
14. J. Pei, J. Han, B. Mortazavi-Asi, H. Pino, "PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth", ICDE'01, 2001.
15. Ayres J., Flannick J., Gehrke J., and Yiu T., "Sequential pattern mining using a bitmap representation", In Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining-2002.

APPLIED ON CLUSTERING ALGORITHM IN EDM

V.Jeyalakshmi, M.Sc., M.Phil.,
Associate Professor, Department of Computer Science
Madurai Sivakasi Nadars Pioneer Meenakshi Women's College, Poovanthi

Abstract

Fifty years ago, there were just a handful of universities across the globe that could provide for specialized educational courses. Today Universities are generating not only graduates but also massive amounts of data from their systems. So, the question that arises is how can a higher educational institution harness the power of this didactic data for its strategic use? This review paper will serve to answer this question. To build an Information system that can learn from the data is a difficult task but it has been achieved successfully by using various data mining approaches like clustering, classification, prediction algorithms etc., however, the use of these algorithms with the educational dataset is quite low. This review paper focuses to consolidate the different types of clustering algorithms as applied in EDM.

Keywords: *clustering, educational data mining (EDM), learning styles, learning management systems (LMS)*

Introduction

According to the international consortium on educational data mining, EDM is defined as a emerging discipline concerned with developing methods for exploring the unique types of data that come from educational settings and using those methods to better understand students and the settings they learn in [1].

EDM focuses on analyzing data generated in an educational setup by the various intra-connected or disparate systems to develop the model for improving the learning experience and institutional effectiveness. Data mining also sometimes referred to as knowledge discovery in databases (KDD) is a known field of study in life sciences and commerce, but the application of data mining to educational context is limited [2].

Various methods have been proposed, applied and tested in data mining field and it is argued by some researchers that these generic methods or algorithms are not suitable to this emerging field of study.

It is proposed that educational data mining methods must be different from the standard data mining methods because of multi-level hierarchy and non-independence in educational data [1]. Institutions are increasing being held accountable for student success [3] since EDM emerged as a sub-discipline in DM there have been notable researches in retention and attrition rates that have been conducted [4]. We applied predictive modeling technique to enhance student retention efforts. Similarly, there have been various software like Weka, Rapid Mineretc.that have been developed to use a combination of DM algorithms or a specific algorithm to aid researchers or stakeholders to find answers to problems but the problem with such tool is that they need to be learned so as to use them. This means that for a novice computer user especially in the administration department of a college or a university, the usage of such tools is not that easy. Commercial e-commerce based websites are using recommender systems that collect user browsing data and recommend similar products there have been efforts to apply the same in the educational context but, they have not been successful as they are highly domain dependent [6].

The objective and purpose of this research paper are to review, different clustering algorithms as applied to EDM context but, with disparate associations. This research paper is to bridge this gap and present a comprehensive review of all types of clustering methodologies as applied to EDM till date.

This paper is organized as follows. Section II is a background of related works about Educational Data Mining (EDM) Section III discusses the various clustering algorithms/techniques applied to an educational dataset. Section IV discusses on the application of clustering algorithms to learning styles of student and learning management systems. Section V provides further discussion and finally Section VI shows the conclusion and future works.

Educational Data Mining

"EDM converts raw data coming from educational systems into useful information. That could potentially have a greater impact on educational research and practice" [7]. Traditionally researchers have applied data mining methods like clustering, classification, association rule mining, text mining to educational context as outlined; [8], conducted a survey that provides a comprehensive resource of papers published between 1995 and 2005 on Educational Data Mining (EDM). Reference [9] has suggested the application of data mining techniques study on-line courses. Had suggested association rules and clustering to support collaborative filtering for the development of e-learning systems. Reference [11] has used a case study that uses prediction methods in a scientific study to game the interactive learning environment by exploiting the properties of the system rather than learning the system.

Reference [12] has provided tools that can be used to support educational data mining. Had shown how educational data mining prediction methods could be used to develop student models. It must be noted that student modeling is an emerging research discipline in educational data mining [1]. While another group of researchers had devised a toolkit that operates within the course management systems and can provide extracted mined information to non-expert users. Data mining techniques have been used to create dynamic learning exercises based on student's progress through a course on English language instruction. While most of the e-learning systems used by educational institutions are used to post or access course materials, they do not provide the educators the necessary tools that could thoroughly track and evaluate all the activities performed by their learners to evaluate the effectiveness of the course and learning process.

Clustering Techniques

The theory of looking at didactic amounts of data whether it's in digital or physical form and stored in diverse repositories be it book keeping records or databases of an educational institution is now termed as big data. According, to Manyika a data set whose computational size exceeds the processing limit of software can be categorized as big data. Several studies have been conducted in the past that have provided detailed insights into the application of traditional data mining algorithms like clustering, prediction, association to tame the sheer voluminous power of big data [9]. Traditional Data Mining algorithms have been applied to various kinds of educational systems. Broadly, the educational system can be classified as two types, brick and mortar based traditional classroom's and the digital virtual classroom's better known as known as LMS System, web-based adaptive hypermedia systems and, intelligent tutoring systems (ITS). The application of various clustering algorithm has been applied in many cases to educational data set in diverse studies. The following table consolidates the research work done on the application of clustering algorithms to an educational dataset.

Using Clustering in EDM

In a learning environment, the learning styles of the student are a decisive factor. In many cases, there has been a mismatch between personal learning styles and the learning demands of different disciplines. Reference, has utilized a two- step cluster analysis approach which examined the brain signals centroids that used electroencephalography (EEG) technology to measure the learning style of participants such that they were successfully able to classify it into four unique clusters. Students typically annotate texts while reading a book by highlighting the context of interest or by underlining it or by writing comments in the side margins. This activity is called annotation.

Researchers have applied statistical clustering method like K-means clustering and Hierarchical clustering to student annotations. And they proved that by using these clustering methods, the creation of students with similar learning style cluster is improved and is faster. Comprehension reading is a very widely used classroom activity in schools and colleges. This helps in building a lifelong reading habit and learning process. This ability of the student behavioral learning patterns has been computationally mapped by applying the Forgy method for k-means clustering and combined with Bloom's taxonomy to determine positive and negative cognitive skills set in about reading comprehension skills. In another study, combined Web-Based Instruction (WBI) programs with the cognitive learning style of the learner to study their effects on student learning patterns. K-means clustering algorithm was used to result in a cluster of students that shared similar learning patterns that further leads to the identification of the related cognitive style for each group.

Learning Management System (LMS) has become an integral part of educational institutions for teaching and learning. Typical LMS logs most of the user activities like course attempted, modules read, practice exam, exam score, student-student interaction via chat logs or discussion boards similarly student-teacher interaction via discussion boards are also logged in the LMS. Several studies have been conducted in this regard.

Reference studied the usage statistics that an LMS provides and, worked on its statistical data analysis and the results were applied in the University of Valencia (Spain). Although they were successful in the statistical analysis of LMS usage data using SPSS but to standardize their methodology the subsequent automation process is yet to be completed and has been left as future work. Performance in exams, usage statistics, regression, number of visits, top search terms, number of downloads of e-learning resources is presented. Several DM approaches and techniques (clustering, classification and association analysis) have been proposed for joint use in the mining of student's assessment data in LMS. Association rules, clustering, classification, sequential pattern analysis, dependency modeling, and prediction have been used to improve web- based learning environments to subsequently enhance the degree to which the educator can evaluate the learning process. Analysis of user access log in Moodle to improve e-learning and to support the analysis of trends is presented in Comparison of different DM algorithms are made to classify learners (predict final marks) based on Moodle usage data. Prediction of student's performance (final grade) based on features extracted from logged data is presented in and, university academic student performance is presented. Prediction of online student’s marks (using an orthogonal search-based rule extraction algorithm) is presented.

Other studies have been conducted to predict student's performance from the log and test scores in web-based instruction (using multi-variable regression), While have used classification, clustering, association rule mining and regression for the discovery of possible dependencies among learner's mean performance and course characteristics. Their results confirm that student’s behavior in an online learning platform affects their performance.

In another study, researchers have shown how educational institutions can benefit from the data collected by LMS. They have proposed an algorithm called "Course Classification Algorithm" when applied in the LMS (Open e-Class platform) that the institution uses can be used to determine and generate course content quality and student online usage reports. These reports are then sent to the instructors for evaluation and motivation purpose. Have proposed the usage of k-means clustering and self- organizing map to cluster learning objects (learning objects are educational resources like eBook, question paper, answer index etc.,) to facilitate faster accessibility of such resources by searching in an LMS. Have proposed Particle Swarm Optimization (PSO)-based clustering for improving the quality of learning by integrating Personalized Learning Environment (PLE) in conjunction with the conventional Learning Management System (LMS).

However, one of the major problems that researchers encounter in finding interesting patterns from educational data set is the relatively small size of the dataset. In another study, we have applied Expectation- Maximization (EM) clustering algorithm to discover student profiles from course evaluation data and for finding associations between subjects that was based on student performance.
Employability of its graduates has been a primary goal of higher educational institutions. Knowledge workers are resorting to key educational courses using Massive Open Online Courses (MOOCs) being provided online by institutions of repute like MIT, Stanford, Harvard to name a few. The year 2012 was witness to rapid development and expansion of several MOOEPs (Massive Open Online Education Platform) like Canvas, Class Togo, Coursera, edX, NPTEL, Udacity to name a few. Had conducted the study to explore the scope of interdisciplinary education through MOOCs. Employability education is an integral component of higher education and an important path by which companies obtain excellent employees. It has been a sustainable argument that in the present socio-economic development, the employability based educational content becomes a mandate.

Discussion

So far we see that subject-specific research has been done but what about domain-specific, i.e., how do institutions employ or apply data mining methods to improve institutional effectiveness? Zimmerman's educational model states that maintaining and monitoring student's academic record is an integral activity of an educational institution. Had used classification algorithm and Prediction algorithm namely decision table and One R algorithm on students" academic record from a previous semester to predict their performance in the current semester. An educational institution maintains and stores various types of student data; it can range from student academic data to their record like parents income, parent's qualification, etc., In a study conducted by they have proved that student's performance can be predicted by using a data set that consisted of student's gender, its parental education, its financial background, etc., Have used Bayesian networks to predict the student outcome based on attributes like attendance, performance in class tests, assignments, etc. Researchers have applied data mining methods like dimensional modeling into educational institutions while others like have used regression analysis and, classification (CS5.0 algorithm which is a type of decision tree) to predict the academic dismissal of students and to predict the GPA of graduated students in e-learning center.

Conclusion and Future Work

The application of data mining methods in the educational sector is an interesting phenomenon. It sets to uncover the previously hidden data to meaningful information that could be used for both

strategic as well as learning gains. In this review paper, we have detailed the various disparate entities that are widely spread across in the educational foray. However, collectively they have not been addressed and this paper serves to bridge this gap. We would continue to pursue our research in clustering algorithms as applied to educational context and will also be working towards generating a unified clustering approach such that it could easily be applied to any educational institutional dataset without any much overhead.

References

1. R. S. J. D. Baker and K. Yacef, "The state of educational data mining in 2009: A review and future visions," Journal of Educational Data Mining, vol. 1, no. 1, 2009.
2. J. Ranjan and K. Malik, "Effective educational process: A datamining approach," Vine, vol. 37, no. 4, pp. 502-515, 2007.
3. B. J. P. Campbell, P. B. Deblois, and D. G. Oblinger, "Academic analytics: A new tool for a new era," Educause Review, vol. 42, pp. 40-57, 2007.
4. J. Luan, Data Mining and Knowledge Management in Higher Education, Toronto, Canada, 2002.
5. S. Lin, "Data mining for student retention management," J. Comput. Sci. Coll., vol. 27, no. 4, pp. 92-99, 2012.
6. O. C. Santos and J. G. Boticario, "Modeling recommendations for the educational domain," Procedia Comput. Sci., vol. 1, no. 2, pp.2793-2800, Jan. 2010.
7. Yan Yan, Qin Xingbin. A Review of Big Data Research in Medicine & Healthcare. e-Science Technology & Application, 2014, 5(6): 3-16.
8. De Oliveira M C F, Levkowitz H. From visual data exploration to visual data mining: a survey [J]. IEEE Transactions on Visualization and Computer Graphics, 2003, 9(3): 378-394.
9. J. R. QUINLAN. Programming for Machine Learning [M]. CA: San Mateo, 1993.
10. Simeon J. Michael H. Arturas Mazeika. Visual Data Mining □An Introduction and Overview□ Springer Berlin/Heidelberg 2008.
11. Bertini E, Lalanne D. Surveying the complementary role of automatic data analysis and visualization in knowledge discovery[C] //Proceedings of the ACM SIGKDD Workshop on Visual Analytics and Knowledge Discovery: Integrating Automated Analysis with Interactive Exploration. New York: ACM Press, 2009: 12-20.
12. J.A. Fails and J. Olsen, "Interactive machine learning," IUI'03: Proceedings of the 8th international conference on intelligent user interfaces, New York, NY, USA: ACM, 2003, pp. 39-45.

UNDERSTANDING THE EFFECTIVENESS OF REGRESSION TESTING TECHNIQUES

S.Priyadharshini
Assistant Professor, Department of Computer Science
Madurai Sivakasi Nadars Pioneer Meenakshi Women's College

Abstract
The most crucial phase of the software development lifecycle is maintenance phase, in which the development team is supposed to maintain the software which is delivered to the clients by them. Software maintenance results for the reasons for error corrections, enhancement of capabilities, deletion of capabilities and optimization. Now the changed or modified software needs testing known as regression testing.
Keywords*: Regression testing, test case prioritization.*

Introduction

Software maintenance is an activity which includes enhancements, error corrections, optimization and deletion of obsolete capabilities. These modifications in the software may cause the software to work incorrectly and may also affect the other parts of the software, so to prevent this Regression testing is performed. Regression testing is used to revalidate the modifications of the software. Regression testing is an expensive process in which test suites are executed ensuring that no new errors have been introduced into previously tested code.

In the second section of this paper I have broadly shown various kinds of regression testing techniques and further discussed classifications of these types given by different authors, then moving into the details of selective and prioritizing test cases for regression testing, consider search algorithms for test case prioritization. In third section I have discussed the methods which may be used to compare various regression testing techniques.

Regression Testing

Regression testing is defined as "the process of retesting the modified parts of the software and ensuring that no new errors have been introduced into previously tested code."

Let P be a program, let P′ be a modified version of P, and let T be a test suite for P. Regression testing consists of reusing T on P′, and determining where the new test cases are needed to actually test code or functionality added to or changed in producing P′.There are various regression testing techniques (1) Retest all; (2) Regression Test Selection; (3) Test Case Prioritization; (4) Hybrid Approach. Figure 1 shows various regression testing techniques.

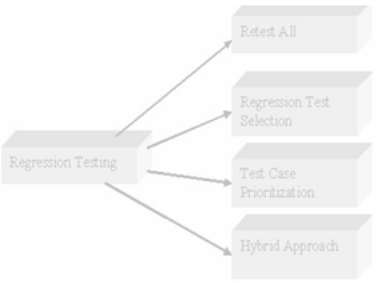


Figure1 Regression Testing Techniques

Retest All

Retest all method is one of the conventional methods for regression testing in which all the tests in the existing test suite are rerun. So the retest all technique is very expensive as compared to techniques

which will be discussed further as regression test suites are costly to execute in full as it requires more time and budget.

Regression Test Selection (RTS)

Due to expensive nature of "retest all" technique, Regression Test Selection is performed. In this technique instead of rerunning the whole test suite we select a part of test suite to rerun if the cost of selecting a part of test suite is less than the cost of running the tests that RTS allows us to omit. RTS divides the existing test suite into (1) Reusable test cases; (2) Retestable test cases; (3) Obsolete test cases. In addition to this classification RTS may create new test cases that test the program for areas which are not covered by the existing test cases. RTS techniques are broadly classified into three categories.

1. Coverage techniques: they take the test coverage criteria into account. They find coverable program parts that have been modified and select test cases that work on these parts.
2. Minimization techniques: they are similar to coverage techniques except that they select the minimum set of test cases.
3. Safe techniques: they do not focus on criteria of coverage; in contrast, they select all those test cases that produce different output with a modified program as compared to its original version.

The various categories in which Regression Test Selection Technique can be evaluated and compared. These categories are: (a) Inclusiveness; (b) Precision; (c) Efficiency; (d) Generality.

a. Inclusiveness is the measure of extent to which a technique chooses the test cases which will cause the changed program to produce different output than the original program, resulting in exposure of faults due to modifications.
b. Precision is the measure of the ability of the technique to prevent choosing test cases that will not make the changed program to produce different output than the original program.
c. Efficiency measures the practicality of a technique.
d. Generality is the measure of ability of a technique to handle complex modifications, realistic language constructs and realistic testing applications.
e. Various techniques of Regression Test Selection as given by various researchers are:

1. Modified noncore function technique: defined in selects test cases that exercise functions in program that have been changed or deleted in producing changed program, or that exercise functions using variables or structures that have been deleted or changed in producing changed program.
2. Modification focused Minimization technique: uses Fischer's approach, which seeks a subset of test suite that is minimal in covering all functions in program identified as changed.
3. Coverage focused Minimization technique: uses the suite reduction technique of Gupta, Harrold, and Soffa to find a subset of test suite that is minimal in covering all functions in program.

Specific methods given by various authors in literature which come under the above mentioned techniques are:

1. Simulating Annealing (SA) algorithm: It is using an optimization formulation of regression testing selection problem.
2. Reduction Methodology (RED): It is a method for reducing the number of selected test cases.

3. Modification Based Reduction version 1 which is an improvement to RED reduces the number of selected regression tests removing tests that cover requirements impacted by the change and those that are redundant.
4. Modification Based Reduction version 2 improves MBR1 by removing tests that cover a requirement.
5. McCabe- based Regression Test Coverage: have got two techniques
   a. Reachability regression Test selection McCabe -based metric (RTM) provides an upper bound of the number of regression tests selected that assures the coverage of requirements affected by the modification at least once.
   b. Data flow Slices regression Test McCabe-based metric (STM) in which we extend McCabe complexity to deal with variable/data modifications. It provides upper and lower bounds to cover the affected definition use pairs created by data modifications.

Test Case Prioritization

This technique of regression testing prioritize the test cases so as to increase a test suite‘s rate of fault detection that is how quickly a test suite detects faults in the modified program to increase reliability. This is of two types: (1) General prioritization which attempts to select an order of the test case that will be effective on average subsequent versions of software.(2)Version Specific prioritization which is concerned with particular version of the software.

Test Case Prioritization Techniques

There are 18 different test case prioritizations techniques numbered P1-P18 which are divided into three groups as shown in figure 2.

Comparator technique

P1: RandomU ordering in which the test cases in test suite are randomly prioritized.
P2: OptimalU orderingU in which the test cases are prioritized to optimize rate of fault detection. As faults are determined by respective test cases and we have programs with known faults, so test cases can be prioritized optimally.

Statement level techniques: (Fine Granularity)

P3: TotalU statement coverage prioritization:U in which test cases are prioritized regarding total number of statements by sorting them in order of coverage achieved. If test cases are having same number of statements they can be ordered pseudo randomly.

P4: AdditionalU statement coverage prioritization:U which is similar to total coverage prioritization, but depends upon feedback about coverage attained to focus on statements not yet covered. This technique greedily selects a test case that has the greatest statement coverage and then iterates until all statements are covered by at least one test case. The moment all statements are covered the remaining test cases undergo Additional statement coverage prioritization by resetting all statements to “not covered”.

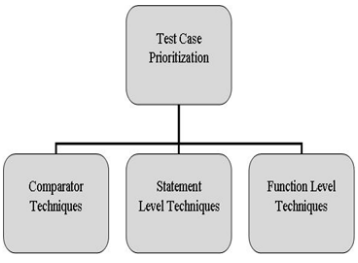


Figure2 Classification of test case prioritization

P5:TotalU FEP prioritization:U in which prioritization is done on the probability of exposing faults

by test cases. Mutation analysis is used to approximate the Fault Exposing Potential (FEP) of a test case. The cost of calculating FEP using mutation analysis is quite high which motivates the search of cost effective approximators of FEP.

P6: AdditionalU FEP prioritization:U the total FEP prioritization is extended to Additional FEP prioritization as the total statement coverage prioritization is extended to Additional statement coverage prioritization.

Function level techniques: (Coarse Granularity)

P7: TotalU function coverage prioritization:U it is similar to total statement coverage but instead of using statements it uses functions. As it has got coarse granularity so the process of collecting function level traces is cheaper than the process of collecting statement level traces in total statement coverage.

P8: AdditionalU function coverage prioritization:U it is similar to Additional statement coverage prioritization with only difference that instead of statements, it is considering function level coverage.

P9: TotalU FEP prioritization (function level):U it is analogous to Total FEP prioritization with only difference that instead of using statements it is using functions.

P10: AdditionalU FEP prioritization (function level):U this technique is similar to Additional FEP prioritization with only difference that instead of using statements it is using functions.

P11: TotalU Fault Index (FI) prioritization:U fault proneness is a measurable software attribute which is used for this technique. Some functions are likely to contain more faults than others, so the fault index is generated using following steps: (1) a set of measurable attributes for each function.(2) the metrics are standardized.(3) principal component analysis which reduces the set of standardized metrics.(4) finally they are combined to a linear function to obtain one fault index per function.

Now for each test case all the fault indexes for every function are added to get total fault index for each test case. Then sort the test cases in decreasing order of these sums to get the result for Total Fault Index (FI) prioritization.

P12: Additional Fault Index (FI) prioritization: as Total function coverage prioritization is extended to Additional function coverage prioritization similarly the Total Fault Index (FI) prioritization is extended to Additional Fault Index (FI) prioritization.

P13: Total FI with FEP coverage prioritization: this technique combines both Total FI and FEP coverage prioritization to achieve a better rate of fault detection.

P14: Additional FI with FEP coverage prioritization: as Total function coverage prioritization is extended to Additional function coverage prioritization similarly Total FI with FEP coverage prioritization is extended to Additional FI with FEP coverage prioritization.

P15: Total Diff prioritization: this technique is similar to Total Fault Index (FI) prioritization with the difference that Total FI prioritization require collection of metrics whereas Total Diff prioritization require only the calculation of syntactic differences between the program and the modified program. Diff means that only syntactic differences are given consideration.

P16: Additional Diff prioritization: Total Diff prioritization is extended to Additional Diff prioritization in a similar way as Total function coverage prioritization is extended to Additional function coverage prioritization.

P17: Total Diff with FEP prioritization: is exactly similar to Total FI with FEP coverage prioritization, except that it is dependent upon changed data derived from diff.

P18: Additional Diff with FEP prioritization: Total Diff with FEP prioritization is extended to Additional Diff with FEP prioritization in a similar way as Total function coverage prioritization is extended to Additional function coverage prioritization.

Search Algorithms for Test Case Prioritization

There are many search techniques for test case prioritization which are being developed and unfolded by various researchers in the field.

1. Greedy algorithm: works on the next best search philosophy. It minimizes the estimated cost to reach a particular goal. Its advantage is that it is cheap in both execution time and implementation. The cost of this prioritization is O (mn) for program containing m statements and test suite containing n test cases.
2. Additional Greedy algorithm: this algorithm [18] uses the feedback from previous selections. It selects the maximum weight element from the part that is not already consumed by previously selected elements. Once the complete coverage is achieved the remaining test cases are prioritized by reapplying the Additional Greedy algorithm. The cost of this prioritization is O (mn^2) for program containing m statements and test suite containing n test cases.
3. 2-Optimal algorithm: Traveling Salesman Problem (TSP) is defined as "find the cycle of minimum cost that visits each of the vertices of a weighted graph G at least once" is solved by K-optimal approach. In case of 2-Optimal algorithm the value of k=2. The cost of this prioritization is O (mn^3) for program containing m statements and test suite containing n test cases.
4. Hill Climbing: it is one of the popular local search algorithms with two variations; steepest ascent and next best ascent. It is very easy and inexpensive to execute. However, this has got cons of dividing O (n^2) neighbors and is unlikely to scale. Steps of algorithm are explained in.
5. Genetic Algorithms (GAs): is a search technique based on the Darwin's theory of survival of the fittest. The population is a set of randomly generated individuals. Each individual is represented by variables or parameters called genes or chromosomes. The basic steps of Genetic Algorithm are (1) Encoding (2) Selection (3) Cross over (4) Mutation.
6. PORT version1.1 (Prioritization of Requirements for Testing) for Test Case Prioritization, which is Requirement Based Test Case Prioritization technique. The technique uses three prioritization factors (1) Customer assigned priority on requirements (2) Requirement Complexity (3) Requirement volatility.

Hybrid Approach

The fourth regression technique is the Hybrid Approach of both Regression Test Selection and Test Case Prioritization. There are number of researchers working on this approach and they have proposed many algorithms for it. For example,

1. Test Selection Algorithm: Implementation of algorithm :
   (a) Input (b) Test Selection algorithm: Adjust module and reduce module (c) output.
2. Hybrid technique combines minimization, modification and prioritization based selection using test history.

Approaches and Challenges

The possible approaches which can be performed so that all the techniques discussed above can be compared and analyzed are: (1) Controlled Experiments (2) Case Studies.

1. Controlled Experiments: are performed on objects drawn from the field and further created and manipulated in controlled environment. The advantage of controlled experiment is that independent variables (eg. test suite, modification patterns, and fault types) can be changed to determine their effect on dependent variables. The disadvantage of this approach is the threat to

external validity shaped by the "manufacturing" of faults, test cases and modifications.

2. Case Studies: are performed on existing programs, several versions and existing test suites. The advantage of this approach is it is "real" and reduces the cost required to be artificially created in Controlled Experiments. The disadvantage of this approach is that certain factors are not controlled, which makes replication difficult.

Thus, each approach –Case Studies and Controlled Experiments have different disadvantages and advantages and comparison and analysis of all the techniques discussed in this paper requires both.

The two major Challenges are: (1) finding objects of study (programs, releases of these programs, test suites) (2) selecting the appropriate approach to answer research questions.

Implementation Result

Regression Testing techniques and have further classified each one of them respectively as explained by various authors, explaining Regression Test Selection and Test Case Prioritization in detail with Search Algorithms for Test Case Prioritization. Through this paper I have tried to, explain the complete structure of Regression Testing, enlighten up the possible areas of Regression Testing.

Conclusion

Regression testing is done in the maintenance phase of the software development life cycle to retest the software for the modifications it has undergone. Approximately 50% of the software cost is involved in the maintenance phase so researchers are working hard to come up with best results by developing new Regression Testing techniques so that the expenditure made in this phase can be reduced to some extent.

References

1. K.K.Aggarwal & Yogesh Singh, "Software Engineering Programs Documentation, Operating Procedures," New Age International Publishers, Revised Second Edition – 2005.
2. Rothermel R., "Efficient Effective Regression Testing Using Safe Test Selection Techniques," Ph.D Thesis, Clemson University, May, 1996.
3. R.A. Johnson and D.W. Wichorn, "Applied Multivariate Analysis," third ed. Englewood Cliffs, N.J.: Prentice Hall, 1992.
4. Hema Srikanth, Laurie Williams, Jason Osborne, "System test case prioritization of new and regression test cases," Proceedings of the seventh international workshop on Economics driven software engineering research, pages 64-73, May .2005.

BIG DATA ANALYTICS TOOLS AND TECHNIQUES

B.Nageswari, M.Sc., M.Phil.,
Assistant Professor of Information Technology
Madurai Sivakasi Nadars Pioneer Meenakshi Women's College, Poovanthi

Abstract
Big data analytics are used to find hidden patterns and consumer preferences for the benefit of organizational.[1]Organizations are capturing, storing, and analyzing data that has high volume, velocity, and variety and comes from a variety of new sources, including social media, machines, log files, video, text, image, RFID, and GPS. The potential value of big data analytics is extreme and is obviously recognized by a growing number of studies.[1] There are keys to success with big data analytics, including a clear business need, strongly committed sponsorship, alignment between the business and IT strategies, as a fact base decision making culture, a well-built data infrastructure, the right analytical tools, and people skilled in the use of analytics. Because of the paradigm shift in the kinds of data being observant and how this data is use, big data can be considered to be a new, 4th generation of decision support data management. Though the business value of big data is vast, especially for online companies like Google and Facebook, how it is being serviced is raising significant privacy apprehension.[2]
Keywords: *Big data, data analysis, tools, applications.*

Introduction

The world today produces the enormous amount of data every day. Intellects have also predicted that this scenario may also result in the great wave of data or dramatically, even a data tsunami. This massive amount of data is nowadays known as Big Data. More or less of the data tsunami being proper, we instantly feel it a necessity to have a tool to have this data in a orderly manner for applications in several disciplines including government, scientific research, industry ,etc . It's will help in a proper study, storage as well as processing of the similar. [1]

Dealing with Information

Let's face it; we want to know things. We're inherently curious. If we recognize the effect of last night's sporting event, we want to know why.

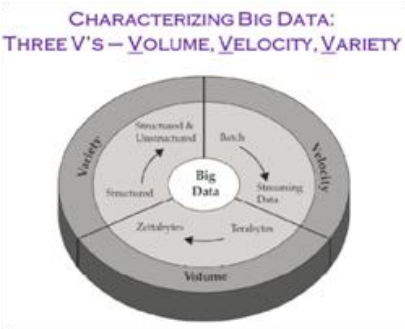
If we incident the temperature now, we want to discern what it will be tomorrow. And if we swatch a magic trick performed before our eyes, we desire to know how it was caused. We can't escape it: it's in our DNA. But as we go through our lives, the questions become more significant and require additional information to resolve. It's getting to the stage where our usual methods are no longer equal to the task. Fortunately, there is an area that concentrates on exactly as these types of problems: it's called Big Data.

What is Big Data?

Big Data is about vast amounts of information. Specifically, it focuses on information sets that are too large to handle in the usual manner. As natural, we mean that they can't be processed by everyday applications, like Microsoft Access or Excel. Unfortunately, even with powerful processors churning away, these applications tend to get bogged down. Add the fact that the size of the information grows each year, and you have a recipe for problems. To get an idea of what we're talking about, consider the amount of raw data the Internal Revenue Service (IRS) processes. It's a wonder we get our tax returns in the time frames we do.

Another useful perspective is to characterize big data as having high volume, high speed, and high variety –.[5]

- High volume – the amount or quantity of data
- High velocity – the rate at which data is created
- High variety – the different types of data



In short, "big data" means there is more of it; it arrives more quickly and totals in more forms. Both of these view are reflected in the following definition: Big data is a term that is used to identify information that is high volume, high velocity, and/or higher variety; requires new technologies and techniques to capture, store, and analyze it; and is used to enhance decision making, provide insight and discovery, and support and optimize processes. [6]

Tools Typically Used In Big Data Scenarios [7]

- NoSQL
- DatabasesMongoDB, CouchDB, Cassandra, Redis, BigTable, HBase, Hypertable, Voldemort, Riak, ZooKeeper
- MapReduce
- Hadoop, Hive, Pig, Cascading, Cascalog, mrjob, Caffeine, S4, MapR, Acunu, Flume, Kafka, Azkaban, Oozie, Greenplum [7]
- Storage
- S3, Hadoop Distributed File System
- Servers
- EC2, Google App Engine, Elastic, Beanstalk, Heroku
- Processing [7]
- R, Yahoo! Pipes, Mechanical Turk, Solr/Lucene, ElasticSearch, Datameer, BigSheets, Tinkerpop[7]

Big Data Analytics

"Big Data" when describing analytics can be misleading. I see "Big Data" as collected raw material. That raw material isn't always a huge amount of data from a single source but can be many different data sets from different resources. The "Internet of Things" would be a great example.[2]

Having this large amount of raw data doesn't offer much value out of the box. However, when applying the correct analytics, we can extrapolate powerful insights.

Some of the common questions that are asked are:

- Where do I begin?
- Are there different types of analytics?
- Which is most impactful for my environment?[2]

Goals of Performing Data Analysis

- You can analyze data.
- Extract actionable and commercially relevant information to boost performance.[8]
- There are a number of extraordinary analytical tools that are free and open source so that you can leverage it to enhance your business and develop skills.

Prescriptive Analytics

Prescriptive analytics is very valuable, but mainly not used. Where big data analytics in general sheds light on a subject, prescriptive analytics gives you a laser-like focus to answer specific questions.[8] For example, in the health care industry, you can better oversee the patient population by using prescriptive analytics to assess the number of patients who are clinically obese, then add filters for factors like diabetes and LDL cholesterol levels to determine where to focus treatment.[9] The same prescriptive model can be applied to almost any industry target group or problem.

Predictive Analytics

Predictive analytics [8] use big data to identify past patterns to predict the time in come. For instance, some societies are using predictive analytics for sales lead scoring. Some companies have gone one step further use predictive analytics for the entire sales process, analyzing lead source, number of communications, types of communications, social media, documents, CRM data, etc. [6] Properly tuned predictive analytics can be applied to sustain sales, marketing, or for other types of complex forecasts.[8]

Diagnostic Analytics

Diagnostic analytics are used for discovery or to determine why something happened. For example, for a social media marketing campaign, you can use descriptive analytics to assess the number of posts, mentions, followers, fans, page views, reviews, pins, etc.[6]

Descriptive Analytics

Descriptive analytics or data mining is at the bottom of the big data value chain, but they can be valuable for uncovering patterns that offer insight. [9] Descriptive analytics can be useful in the sales cycle, for example, to categorize customers by their likely product preferences and sales cycle.[8]

Conclusion

Handling big data efficiently is the need of the hour and one needs to come up with plausible solutions to these challenges one needs to understand the concept of big data, its handling methodologies and furthermore improve the approaches in analyzing big data. With the advent of social media the need for handling big data has increased monumentally. As more and more organizations are stepping out of the traditional boundaries big data keeps growing bigger.[10] The tools being developed are efforts for overcoming the challenges arising due to big data.

References

1. http://searchbusinessanalytics.techtarget.com/definition/big-data-analytics
2. https://dzone.com/articles/big-data-analytics-delivering-business-value-at-am
3. https://arxiv.org/ftp/arxiv/papers/1705/1705.04928.pdf
4. https://www.sas.com/en_us/insights/analytics/big-data-analytics.html
5. http://whatis.techtarget.com/definition/3Vs
6. https://www.quora.com/What-is-the-best-training-institute-for-Big-Data-in-Chennai?no_redirect=1
7. https://www.slideshare.net/kundankumar18847876/introduction-to-big-data-and-hadoop
8. https://underworldtricks.com/data-analytics-software/
9. http://files.ebook777.com/015/Big%20Data%20Analytics%20with%20R%20and%20Hadoop.pdf
10. http://cdn.oreillystatic.com/oreilly/radarreport/0636920028307/Big_Data_Now_2012_Edition.pf

APPLICATIONS OF GENETIC ALGORITHMS IN DATA MINING

K.Sudharani
Head of the Department, Department of Computer Science
Madurai Sivakasi Nadars Pioneer Meenakshi Women's College, Poovanthi

Abstract
Data mining is the mining, or discovery, of new information in terms of patterns or rules from vast amount of data. To be useful, data mining must be carried out efficiently on large files and databases. To extract the knowledge, a database may be considered as a large search space and a mining algorithm as a search strategy. Genetic Algorithm (GA) is a search-based optimization technique based on the principles of Genetics and Natural Selection. It is frequently used to find optimal or near-optimal solutions to difficult problems which otherwise would take a lifetime to solve. It is frequently used to solve optimization problems, in research, and in machine learning. Genetic algorithms are commonly used to generate high-quality solutions to optimization and search problems by relying on bio-inspired operators such as mutation, crossover, and selection. GA evolves towards better solutions over some generations. Search strategies based on genetic-based algorithms have been applied successfully in a wide range of applications. In this paper, the suitability of genetic-based algorithms for data mining is discussed. This paper focuses on the various application areas where genetic algorithm plays an evolutionary role with data mining technique and explains them in details.
Keywords*: Data Mining, Genetic Algorithm, Prediction, Classification.*

Introduction

Genetic Algorithms were invented to mimic some of the processes observed in natural evolution. GA is simulations of evolution, of what kind ever. In most cases, however, genetic algorithms are nothing else than probabilistic optimization methods which are based on the principles of evolution.

Genetic Algorithms (GAs) are adaptive heuristic search algorithm based on the evolutionary ideas of natural selection and genetics. The basic techniques of the GAs are designed to simulate processes in natural systems necessary for evolution; especially those follow the principles first laid down by Charles Darwin of "survival of the fittest." Since in nature, competition among individuals for small resources results in the fittest individuals dominating over the weaker ones.

The set of all the solutions of an optimization problem constitutes the search space. The problem consists in finding out the solution that fits the best, from all the possible solutions. When the search space becomes huge, we need a specific technique to find the optimal solution. GAs provides one of these methods. Practically they all work in a similar way, adapting the simple genetics to algorithmic mechanisms. GA handles a population of possible solutions. Each solution is represented through a chromosome, which is just an abstract representation.

Data Mining

Data mining is a promising computational paradigm that enhances traditional approaches to discover and increase the opportunities for understanding complex physical and biological systems. It is a decision support process in which there is search for patterns of information in data. Data mining uses sophisticated statistical analysis and relationships hidden in organizational databases. Once the patterns are found the information is to be presented in a suitable form with graphs, reports etc. Two critical factors for success with data mining are a large well integrated data warehouse and a well-defined understanding of the process within which data mining is to be applied. Data mining involves

the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets. These tools can include statistical models, mathematical algorithms, and machine learning methods (algorithms that improve their performance automatically through experience, such as neural networks or decision trees). Consequently, data mining consists of more than collecting and managing data, it also includes analysis and prediction.

A number of advances in technology and business processes have contributed to a growing interest in data mining in both the public and private sectors. Some of these changes include the growth of computer networks, which can be used to connect databases; the development of enhanced search-related techniques such as neural networks and advanced algorithms; the spread of the client/server computing model, allowing users to access centralized data resources from the desktop; and an increased ability to combine data from disparate sources into a single search source.

Data

Data are any facts, numbers, or text that can be processed by a computer. Today, organizations are accumulating vast and growing amounts of data in different formats and different databases. This includes:

- Operational data or transactional data such as sales, cost, inventory, payroll, and accounting
- Non-operational data, such as industry sales, forecast data, and macroeconomic data
- Metadata - data about the data itself, such as logical database design or data dictionary definitions

Information

The patterns, associations, or relationships among all this data can provide information.

Knowledge

The knowledge base is helpful in the whole data mining process. It might be useful for guiding the search or evaluating the interestingness of the result patterns. The knowledge base might even contain user beliefs and data from user experiences that can be useful in the process of data mining. The data mining engine might get inputs from the knowledge base to make the result more accurate and reliable. The pattern evaluation module interacts with the knowledge base on a regular basis to get inputs and also to update it.

Data Warehouse

Dramatic advances in data capture, processing power, data transmission, and storage capabilities are enabling organizations to integrate their various databases into data warehouses. Data warehousing is defined as a process of centralized data management and retrieval. Data warehousing, like data mining, is a relatively new term although the concept itself has been around for years. Data warehousing represents an ideal vision of maintaining a central repository of all organizational data. Centralization of data is needed to maximize user access and analysis. Dramatic technological advances are making this vision a reality for many companies. And, equally dramatic advances in data analysis software are allowing users to access this data freely. The data analysis software is what supports data mining.

How does data mining work?

Data mining is the process of exploration and analysis, by automatic or semi automatic means, of large quantities of data in order to discover meaningful patterns and rules. They scour databases for hidden patterns, and find predictive information that experts may miss as it lies outside their

expectations. Data mining functionalities are used to specify the kinds of patterns to be found in data mining tasks. In general, such tasks can be classified into two categories: predictive and descriptive. Predictive mining tasks perform induction on the current data in order to make predictions. Predictive modeling may be made based on the use of other historical data. Predictive model data mining tasks include classification, regression, time series analysis, and prediction.

A descriptive mining task identifies patterns or relationships in data. A descriptive model serves as a way to explore the properties of the data examined, not to predict new properties. Clustering, Summarization, association rules, and sequence discovery are usually viewed as descriptive in nature.

1. Classification: Classification maps data into predefined groups or classes. It is often referred to as supervised learning because the classes are determined before examining the data.
2. Regression: Regression is used to map a data item to a real valued prediction variable. In actuality, regression involves the learning of the function that does this mapping. Regression assumes that the target data fit into some known type of function and then determines the best function of this type that models the given data. Some type of error analysis is used to determine which function is "best"
3. Time Series Analysis: With time series analysis, the value of an attribute is examined as it varies over time. The values usually are obtained as evenly spaced time points (daily, weekly, hourly etc.). A time series plot is used to visualize the time series.
4. Prediction: Prediction can be viewed as a type of classification. The difference is that prediction is predicting a future state rather than a current state.
5. Clustering: Clustering is similar to classification except that the groups are not predefined, but rather defined by the data alone. Clustering is alternatively referred to as unsupervised learning or segmentation.
6. Summarization: Summarization maps data into subsets with associated simple descriptions. Summarization is also called characterization or generalization. It extracts or derives representative information about the database. This may be accomplished by actually retrieving portions of the data.
7. Association rules: Link analysis, alternatively referred to as affinity analysis or association, refers to the data mining task of uncovering relationships among data.
8. Sequence discovery: Sequential analysis or sequence discovery is used to determine sequential patterns in data. These patterns are based on a time sequence of actions. These patterns are similar to associations in that data are found to be related, but the relationships are based on time.

Data mining consists of five major elements

- Extract, transform and load transaction data onto the data warehouse system.
- Store and manage the data in a multidimensional database system.
- Provide data access to business analysts and information technology professionals.
- Analyze the data by application software.
- Present the data in a useful format, such as a graph or table.

Use of Genetic Algorithm in Data Mining

Data mining algorithms require a technique that partitions the domain values of an attribute in a limited set of ranges, because considering all possible range of domain values is infeasible.

Applications of Genetic Algorithm

Genetic Algorithms are primarily used in optimization problems of various kinds, but they are frequently used in other application areas as well. A heuristic search technique used in computing and Artificial Intelligence to find optimized solutions to search problems using techniques inspired by evolutionary biology: mutation, selection, reproduction [inheritance] and recombination.

Optimization – Genetic Algorithms are most commonly used in optimization problems wherein we have to maximize or minimize a given objective function value under a given set of constraints.

Neural Networks – GAs are also used to train neural networks, particularly recurrent neural networks.

Parallelization – GAs also have very good parallel capabilities, and prove to be very effective means in solving certain problems, and also provide a good area for research.

Image Processing – GAs are used for various digital image processing (DIP) tasks as well like dense pixel matching.

Scheduling applications – GAs are used to solve various scheduling problems as well, particularly the time tabling problem.

Robot Trajectory Generation – GAs have been used to plan the path which a robot arm takes by moving from one point to another.

Parametric Design of Aircraft – GAs have been used to design aircrafts by varying the parameters and evolving better solutions.

DNA Analysis – GAs have been used to determine the structure of DNA using spectrometric data about the sample.

Traveling salesman problem and its applications – GAs have been used to solve the TSP, which is a well-known combinatorial problem using novel crossover and packing strategies.

Marketing and Merchandising - GAs are indeed being put to work to help merchandisers to produce products and marketing consultants design advertising and direct solicitation campaigns to sell stuff. Maybe this application of GAs could someday get us out of the financial black hole and get things moving again.

Literature Survey

Educational Data Mining is concerned with developing methods that extract knowledge from data originating from educational domain. The extracted information that describes student performance can be stored as knowledge for decision making to improve the quality of education in institutions. Educational Data mining (EDM) has become a very promising research area [1]. The Educational Data Mining (EDM) focuses on modeling and evaluates student's performance based on examination scores and college atmosphere evaluation questionnaire. In the academic information system, the data mining is mainly used for predicting the students' performance [2]. During the credit card transaction, the fraud is detected and the number of false alert is being minimized by using genetic algorithm. The high amount of losses due to fraud and the awareness of the relation between loss and the available limit have to be reduced by using genetic algorithm [3]. By embedding the frequent schemata into the GA evolution process, the new improved GA could reduce the search time by preserving segments of better solutions without accidentally being lost due to random crossover or mutation. The proposed new GA experimented on a 6x6 job shop scheduling problem [4]. A detailed Study conducted on the DJIA stock market and the various technical indicators used for Stock Market data which help in analyzing the Stock Market. The Literature survey conducted on Genetic Algorithm, and Association Rule Mining Algorithm and a combined approach were being implemented. As Association Rule Mining Algorithms

cannot handle numerical data efficiently, a modified Genetic Algorithm was being used for the representation of the stock market data and to generate rules among the various technical indicators [5].

Conclusion

Genetic algorithms provide a comprehensive search methodology for machine learning and optimization. It has been proved to be efficient and powerful through application of data mining techniques that use optimization and classification. GAs can rapidly locate solutions, in data mining even for difficult search spaces. GAs are used in various fields of Data mining to get the optimized solutions for the better performance of the data that are essential in decision making and process the accurate result. In this paper, integration of the genetic algorithm with data mining was discussed to obtain optimized results for better performance of the data which are required in decision making and process the accurate result. Thus genetic algorithms can be used in real analysis systems to achieve a better solution.

References

1. R.Sumitha, E.S.Vinothkumar, "Prediction of Students Outcome Using Data Mining Techniques," International Journal of Scientific Engineering and Applied Science (IJSEAS) – Volume-2, Issue - 6, June 2016 ISSN: 2395-3470
2. K.Lakshmipriya, Dr. P.K. Arunesh, "Predicting Student Performance Using Data Mining Classification Techniques," International Conference on Recent Trends in Engineering Science, Humanities, and Management ISBN: 978-93-86171-18-4
3. K.RamaKalyani, D.UmaDevi, "Fraud Detection of Credit Card Payment System by Genetic Algorithm," International Journal of Scientific & Engineering Research Volume 3, Issue 7, July-2012 ISSN 2229-5518
4. S. Wesley Changchien & Ya-Tai Lin, "Use Data Mining to Improve Genetic Algorithm Efficiency for a Job-shop Scheduling Problem," The First International Conference on Electronic Business, Hong Kong, December 19-21, 2001.
5. Shruti Samant, "Prediction of Financial Performance Using Genetic Algorithm and Associative Rule Mining," International Journal of Engineering Research and General Science Volume 3, Issue 1, January-February, 2015 ISSN 2091-2730
6. Gunjan Verma, Vineeta Verma, " Role and Applications of Genetic Algorithm in Data Mining," International Journal of Computer Applications (0975 – 888) Volume 48– No.17, June 2012
7. M.Srinivas, L.M.Patnaik, "Genetic Algorithms: a survey," International Journal of Computer Applications (17-26) ISSN: 0018-9162
8. John McCall, "Genetic Algorithms for modeling and optimization," Journal of Computational and Applied Mathematics (205 - 222), Volume 184, Issue 1, December 2005.
9. Saad Alharbi, Ibrahim Venkat, "A Genetic Algorithm Based Approach for Solving Minimum Dominating Set of Queens Problem," Journal of Optimization Volume 2017 (2017), Article ID 5650364, eight pages.
10. David E.Goldberg, Chie Hsiung Kuo, "Genetic Algorithms in Pipeline Optimization," Journal of Computing in Civil Engineering, (ASCE)0887-3801(1987)
11. Bin Yu Zhongzhen Yang, Jinbao Yao, "Genetic Algorithm for Bus Frequency Optimization," Journal of Transportation Engineering / Volume 136 Issue 6 – June 2010
12. He'lio Fiori de Castro, Katia Lucchesi Cavalca, "Availability optimization with genetic algorithm," International Journal of Quality & Reliability Management ISSN: 0265 – 671X.

A REVIEW ON BIG DATA IN CLOUD COMPUTING

A.Saranya, M.Sc., M.Phil.,
Assistant Professor, Department of Information Technology
Madurai Sivakasi Nadars Pioneer Meenakshi Women's College, Poovanthi

Abstract
In this paper introduces several big data and cloud computing techniques which can store and process the varied volumes of data, providing both enterprises and science with insights over its clients/experiments. In this paper we present an overview of both technologies and cases of success when integrating big data and cloud frameworks. Hence, big data solves much of our current problems it still gives some gaps and issues that raise concern and need more improvement. Security, privacy, scalability, data governance policies, data heterogeneity, disaster recovery mechanisms, and other challenges are yet to be discussed.
Keywords: *Big data, Cloud Computing, Data Management, Big data Issues*

Introduction

Nowadays, there has been an increasing demand to store and process more and more data, in domains such as finance, science, and government. Systems that support big data, and host them using cloud computing, have been developed and used successfully.

Big data is initiated from the web companies who used to handle loosely structured or unstructured data. "Each day, we create 2.5 quintillion bytes of data so much that 90% of the data in the world today has been created in the last two years alone. Lots of data is being collected and warehoused like,

- Web data, e-commerce
- Bank/Credit card transactions
- Social Network

Society is becoming more instrumented and as a result, organizations are producing and storing vast amounts of data. Data is becoming more valuable[1].

Big data technology has been gaining popularity in severaldomains of business, engineering, and scientific computing areas; Big data is referred to as a collection of datasets having vast amount of data ranging in zettabytes [2]. Big Data is becoming linked with almost all aspects of human activity from a research point of view to digital services. Latest technologies such Hadoop, HDFS, MapReduce and network connectivity provide an interface for automation of all processes like a collection of datasets, storage, processing and visualization [3, 4, 5, 6].Bigdata is the burning topic of research and researchers has a great urge to dig more and more about it and generate best possible answers to the questions still unknown.

Big data and Cloud Computing

The concept of big data became a important force of innovation across both academics and corporations. The architype is viewed as an effort to understand and get proper insights from big datasets (big data analytics), providing concise information over huge data loads.

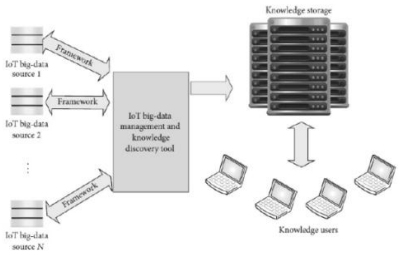


Fig.1 Bigdata Knowledge Discovery

Though big data is mostly deals with the storage of huge loads of data it also concerns ways to process and extract knowledge from it [7]. The five different aspects used to describe big data (commonly referred to as the five "V"s) are Volume, Variety, Velocity, Value and, Veracity [8]:

Dimensions of Big Data There are many properties associated with big data. The prominent aspects are volume, variety, velocity, variability and value.

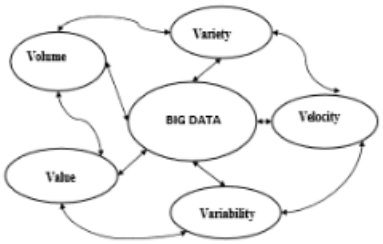


Fig.2 Five V's of big data

Volume: many factors contribute to the increase in volume like storage of data, live streaming, etcdescribes the size of datasets that a big data system deals with Processing and storing big volumes of data is rather difficult, since it concerns: scalability so that the system can grow; availability, which guarantees access to data and ways to perform operations over it; and bandwidth and performance.

Variety: various types of data is to be supported. *It* concerns the different types of data from various sources that big data frameworks have to deal with.

Velocity: The speeds at which the files are created and processed are carried out refers to the velocity. It concerns the different rates at which data streams may get in or out the system and provides an abstraction layer so that big data systems can store data independently of the incoming or outgoing rate.

Variability: it describes the amount of variance used in summaries kept within the data bank and refers how they are spread out or closely clustered within the data set. Also refers to the trustworthiness of the data, addressing data confidentiality, integrity, and availability.

Value: all enterprises and e-commerce systems are keen in improving the customer relationship by providing value added services and concerns the true value of data (i.e., the potential value of the data regarding the information they contain). Huge amounts of data are worthless unless they provide value.

Cloud Computing

Cloud computing is another paradigm which promises theoretically unlimited on-demand services to its users. Cloud's ability to virtualize resources allows abstracting hardware, requiring little interaction with cloud service providers and enabling users to access terabytes of storage, high processing power, and high availability in a *pay-as-you-go* model.

The term Cloud refers to a Network or Internet. In other words, we can say that Cloud is something, which is present at the remote location. Cloud can provide services over a network, i.e., on public networks or private networks, i.e., WAN, LAN or VPN. Applications such as e-mail, web conferencing, customer relationship management (CRM), all run in the cloud.

Cloud computing is a distributed architecture with centralizing server. The Cloud computing is depends on internet technology which provides computing services in the form of Infrastructure as services (Iaas), platforms as service (Paas) and Software as Service (SaaS) to the user. The user does not require knowledge or expertise to control the infrastructure of clouds; it provides only the abstraction.

It can be utilized as a service of an Internet with high scalability, higher throughput, quality of service and high computing power.

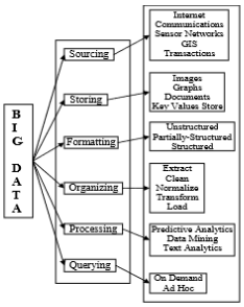


Fig.3 Five V's of big data

Infrastructure as a Service (IaaS): The IaaS layer offers storage and infrastructure resources that areneeded to deliver the Cloud services. It only comprises of the infrastructure or physical resource.

Platform as a Service (PaaS): PaaS provides the combination of both, infrastructure and application. Hence, organizations using PaaS don't have to worry for infrastructure nor services.

Software as a Service (SaaS): In the SaaS layer, the Cloud service provider hosts the software on their servers. It can be defined as amodel in which applications and software's are hosted on the server and made available to customers over a network.

Cloud Deployment Models

Cloud analytics solutions need to consider the multiple Cloud deployment models adopted by enterprises are:

Private*:* These mainly work on the private network, managed by the organisation itself or by the third party. A private Cloud is suitable for businesses that require the highest level of control of security and data privacy.

Public*:* These work with off-site over the Internet and available to the general public. Public Cloud offers high efficiency and shared resources with low cost. The quality of services such as Privacy, security, and availability is specified in a contract.

Hybrid: combines both Clouds where additional resources from a public Cloud can be provided as needed to a private Cloud. Customers can develop and deploy analytics applications using a private environment.

Community Cloud: This cloud infrastructure is shared by several organizations.

Big Data in Cloud

Storing and processing big volumes of data requires scalability, fault tolerance and availability. Cloud computing delivers all these through hardware virtualizations. Thus, big data and cloud computing are two compatible concepts as cloud enables big data to be available, scalable and fault tolerant. Business opportunity. As such, several new companies such as Cloudera, Teradata and many others, have started to focus on delivering Big Data as a service(BDaaS) or database as a service(DBaaS). Companies such as Google, IBM, Amazon and, Microsoft also provide ways for consumers to consume big data on demand. Next, we present to examples, Nokia and RedBus, which discuss the successful use of bigdata within Cloud environment.

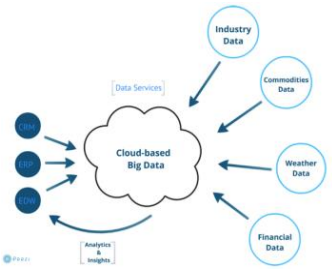


Fig.4 Big data in cloud

Cloud comes with an explicit security challenge, that is the data owner might not have any control of where the data is placed. The reason behind this control issue is that if one wants to get the benefits of cloud computing, he/she must also utilize the allocation of resources and also the scheduling given by the controls. Hence it is required to product the data in the midst of untrustworthy processes.

Cloud Computing Challenges

Security and Privacy

Security and Privacy of information is the biggest challenge to cloud computing. Security andprivacy issues can be overcome by employing encryption, security hardwareand securityapplications.

Portability

This is another challenge to cloud computing that applications should easily be migrated from singlecloud provider to another. There must not be vendor lock-in. However, it is not yet made possiblesince each of the cloud providers uses different standard languages for their platforms.

Interoperability

It means the application on one platform should be able to incorporate services from the extraplatforms. It is made possible via web services, but developing such web services is very complex.

Computing Performance

Data-intensive applications on cloud require high network bandwidth, which results in high cost.Low bandwidth does not meet the desired computing performance of cloud application.

Reliability and Availability

It is necessary for cloud systems to be reliable and robust because most of the businesses are nowbecoming dependent on services provided by third-party.

Cloud Issues

Data Security: Data Security refers as confidentiality, integrity and, availability. These are the major issues for cloud vendors. Confidentiality is defined to a privacy of the user data in the cloud system.

Data Locations: When users use, they probably won't know exactly where their data will have hosted and which location it will be stored in. In fact, they might not even know what country it will be stored in. [9]

Trust Issue: Trust is also a major issue in cloud computing. Trust can be in between human to machine, machine to human, human to human, machine to human. In cloud computing, the user stores their data on cloud storage because of trust on the cloud.

Data Recovery: It is defined as the process of restoring data that has been lost, corrupted or accident.

Big Data Issues and Challenges

Issues	Existent solutions	Advantages	Disadvantages
Security	Based on SLAs and Data Encryption	Data is encrypted	Querying encrypted data is time-consuming
Privacy	-De-identification -User Consent	Provides a reasonable privacy or transfers responsibility to the user	It was proved that most de-identification mechanisms colud be reverse engineered
Heterogeneity	One of the big data systems' characteristics is the ability to deal with different data coming at different velocities	The major types of data are covered up	It is difficult to handle such variety of data and such different velocities

Data Governance	Data governance documents	-Specify the way data is handled; -Specify data access policies; -Role specification; -Specify data life cycle	-The data life cycle is not easy to define; -Enforcing data governance policies so much can lead to counterproductive effects
Disaster recovery	Recovery plans	Specify the data recovery locations and procedures	Normally there is only one destination from which to secure data
Data Uploading	-Send HDDs to the cloud provider -Upload data through the Internet	Physically sending the data to the cloud provider is quicker than uploading data but it is much more unsecure	Physically sending data to the cloud provider is dangerous as HDDs can suffer damage from the trip. - Uploading data through the network is time-consuming and, without encryption, can be insecure
High Data processing (Exabyte datasets)	-Cloud computing - HPCs	Cloud computing is not so cost expensive as HPCs but HPCs are believed to handle Exabyte datasets much better	HPCs are very much expensive ant its total cost over a year is hard to maintain. On the other hand, cloud is believed that cannot cope with the requirements for such huge datasets
Scalability	Scalability exists at the three levels in the cloud stack. At the Platform level there is: horizontal (Sharding) and vertical scalability	Scalability allows the system to grow on demand	Scalability is mainly manual and is very much static. Most big data systems must be elastic to cope with data changes
Elasticity	There are several elasticity techniques such as Live Migration, Replication and Resizing	Elasticity brings the system the capability of accommodating data peaks	Most load variations assessments are manually made, instead of automatized

Conclusions

With data increasing on a daily base, big data systems and in particular, analytic tools, have become a major force of innovation that provides a way to store, process and get information over petabyte datasets. Cloud environments strongly leverage big data solutions by providing fault-tolerant, scalable and available environments to big data systems.

References

1. Marcos D. Assunção, Rodrigo N. Calheiros, Silvia Bianchi, Marco A.S. Netto, Rajkumar Buyya “Big Data computing and clouds: Trends and future directions”, journal of parallel and distributed Computing. Aug 25 2014.
2. Pal, A., & Agrawal, S. (2014, August). An experimental approach towards big data for analyzing memory utilization on a hadoop cluster using HDFS and Map Reduce. In Networks & Soft Computing (ICNSC), 2014 First International Conference on (pp. 442-447). IEEE.
3. Bedi, P., Jindal, V., & Gautam, A. (2014, September). Beginning with big data simplified. In Data Mining and Intelligent Computing (ICDMIC), 2014 International Conference on (pp. 1-7). IEEE.

4. Demchenko, Y., De Laat, C., & Membrey, P. (2014, May). Defining architecture components of the Big Data Ecosystem. In Collaboration Technologies and Systems (CTS), 2014 International Conference on (pp. 104-112). IEEE.
5. Demchenko, Y., Grosso, P., De Laat, C., &Membrey, P. (2013, May). Addressing big data issues in scientific data infrastructure. In Collaboration Technologies and Systems (CTS), 2013 International Conference on (pp. 48-55). IEEE.
6. Wood, J., Andersson, T., Bachem, A., Best, C., Genova, F., Lopez, D. R...& Vigen, J. (2010). Riding the wave: How Europe can gain from the rising tide of scientific data. Final report of the High Level Expert Group on Scientific Data; A submission to the European Commission, October 2010.
7. Hashem, I.A.T. et al., 2014. The rise of "big data" on cloud computing: Review and open research issues. *Information Systems*, 47, pp.98–115.
8. Sakr, S. & Gaber, M.M., 2014. *Large Scale and big data: Processing and Management* Auerbach, ed.,
9. Feng-Tse Lin, Teng-San Shih, "Cloud Computing: The Emerging Computing Technology," ICIC Express Letters Part B: Applications (ISSN: 2185-2766), v1, September 2010, pp. 33-38.

SEMANTIC METHODOLOGY FOR SEMI-AUTOMATIC ONTOLOGY CONSTRUCTION USING ONTOLOGY LEARNING – COMPARATIVE STUDY OF ALGORITHMS

B.Gomathi, MCA., M.Phil.,
Assistant Professor, Department of Computer Science
Madurai Sivakasi Nadars Pioneer Meenakshi Women's College, Poovanthi

Abstract
The analysis of large amounts of machine-generated data requires innovative methods. We propose the combination of Semantic Web with ontology construction and Machine Learning to facilitate this survey. First, collected data is preprocessed and converted to RDF according to a schema in the Web Ontology Language OWL. After that they are classified and tested with supervised, unsupervised and semi-supervised algorithms. We found that supervised learning algorithms give the better approach than the other algorithms based on performance metrics.
Keywords - *Machine learning, RDF, OWL, semantic web, ontology, semantic web;*

Introduction

Machine learning is a field of computer science that manage computers with the ability to learn without being explicitly programmed. Machine learning will continue to develop the innovation with algorithms that help computers learn and refine responses and adapt to new data and content input increasingly. Machine learning models, together with semantic enrichment and ontology construction are capable of classifying and clustering content so that it can be easily reused for anytime and repackaged for other models. This approach enables publishers to offer highly-relevant personalized adaptive content to engage, retain and win over readers, educators, and news consumers. The Modern information system is moving from data processing towards concept processing.

Semantic web Technologies offer a new approach to manage the user defined information and processes by adding special meaning to the data. Ontology is a structure oriented concept and the knowledge about a particular domain is described with relevant concepts and relations between the words [1]. This semantically-enriched approach reduces production costs via content re-use and re-purposing. Ontologies act as a shared vocabulary for assigning data semantics. To the extent that RDF-based techniques are useful as data-handling technologies, they ought to be entirely complementary to the recent breakthroughs achieved via deep learning. Semantic Web can provide formats for data and metadata that is processed under ML systems (just as it can do the same for systems that are built from Perl, XSLT, SQL or that even have humans in the loop). Ontology construction was done with the Protégé environment with the usage of OWL language [6].

The remainder of the paper is organized as given below: Section 3 presents current knowledge management techniques; Section 4 presents a summary of research work and, future scope.

Related Work

In the last few years, there has been an increasing interest in the application of machine learning (ML) [2]. The machine learning community was one of the first to employ representations that incorporated probabilities. It is clear that we need to find ways of representing expectations in the Semantic Web community as well. Possible models for RDF face the challenge that real- world triples are too small and ambiguous a data structure to use for probabilities. Ontology learning is a formal

representation of conceptual knowledge. It consists of different components of concepts, relationships, instances, meaning and axioms about classes and properties. The OWL language provides the mechanism for creating all the elements, which is needed for on ontology. There are two types of properties as object properties and data properties [4].

The limited number of allowed syntactic sentence structure, make the language easier to learn. It is easier to use than the understanding OWL, RDF or SQL [7]. Supervised, semi – supervised and unsupervised algorithms are compared with different algorithms [8]. As the proof-of-concept, we present a preliminary implementation of the architecture and apply it to several variants of a simple video game [9]. For the acquisition of taxonomic relations, we exploit inherent multi-word terms' lexical information in a comparative implementation of agglomerative hierarchical clustering and formal concept analysis methods. For the detection of non-taxonomic relations, we comparatively investigate in Onto Gain an association rule based algorithm and a probabilistic algorithm.[10]. K-means algorithm is characterized by fastness of clustering and easiness of realization. It is a typical distance-based clustering algorithm, adopting distance as the evaluating indicator of I similarity, i.e., the different is the distance between two objects, and the more similar they are. The final goal is to obtain a cluster which is tight and independent[11].

Methodology for Machine Learning

Machine learning tasks are typically classified into two main categories, depending on whether there is a learning "signal" or "feedback" available to a learning system, they falls under three categories:

1. Supervised learning
2. Unsupervised learning
3. Semi- supervised learning

Ontology Learning

Ontology learning includes learning ontology concepts based on word's meaning, learning ontology relationships between existing concepts, learning both the methods and relations at a time, populating the existing ontology structure, and dealing with dynamic data streams. It also includes construction of ontologies giving different views on the same data. It is for finding the mapping among ontology components. Data on the web is of heterogeneous type. It consists of text documents, images, data records, etc. Another problem of web data is of dynamic nature, as the data on the web is continuously updated. By applying unsupervised learning algorithm such as clustering, the similarity between the objects used within documents can be identified. In semiautomatic approach, ontology is constructed from the titles(concepts) from the documents data. The Machine can provide suggestions for the topics appearing and can assists human by automatically assigning documents to the topics. It suggest the naming of the topics. From the set of documents, possible concepts are retrieved and relationships are identified with applying the document clustering and classification or similarity measure based on semantics [1]. Latent Semantic Indexing is based on the concept of finding similarity words with similar meanings.

Fig.1 Ontology construction

Supervised learning

The computer is presented with example inputs and their desired outputs, given by the "admin", and the goal is to learn a general rule and methods that maps inputs to outputs. As in many cases, the input signal can be only partially available, or restricted to special feedback.

Algorithms: K-means clustering algorithms

Basically, K-Means runs on distance calculations, which again uses “Euclidean Distance” for this purpose. Euclidean distance calculates the distance between two given points using the following formula:

EuclideanDistance = $\sqrt{(X_1 - X_2)^2 + (Y_1 - Y_2)^2} \rightarrow (1)$

Semi-supervised learning

The computer give only an incomplete training signal: a training set with some (often many) of the target outputs missing.

Algorithm: Graph Laplacian

Let $0 < \chi, \gamma, \beta, \alpha < 1$ be real numbers and =1. The similarity function can be$\chi + \gamma + \beta + \alpha$ represented as the weighting sum of their name similarity, structure similarity, instance similarly and attribute similarity:

$Simstructe(A,B)+\beta Simname(A,B) +\alpha Sim(A,B) = Simattribute(A,B),\chi Siminstance(A,B)+ \gamma$ --(2)

Active learning

The computer can only obtain training labels data for a limited set of instances (based on a budget), and also has to find its choice of objects to acquire labels for. When used interactively, these can be presented to the user for labeling.

Algorithm: Weighing schema algorithms (Naïve Bayes)

To extract concept relations from unstructured text using a syntactic and semantic probability-based Naïve Bayes classifier. We propose an algorithm to iteratively extract a list of attributes, and associations for the seed concept from which the rough schema is conceptualized

Reinforcement Learning

Training data (in the form of rewards and punishments) is given only as feedback to the program's actions in a dynamic environment, such as driving a vehicle or playing a game against an opponent.

Algorithms: Hybrid neural-symbolic, end-to-end reinforcement

Integrating the soft retrieval process with a reinforcement learner leads to higher task success rate and reward in both simulations and against real users. The fully neural end-to-end agent trained entirely for user feedback, and discusses its application towards personalized dialogue agents.

Unsupervised Learning

No labels are given to the learning algorithm, leaving it on its own to find structure in its input. Unsupervised learning can be a goal in itself (discovering hidden patterns in data) or, a means towards an end (feature learning).

Algorithms

Unsupervised learning problems can be further grouped into clustering and association problems.

Clustering

A clustering problem is where you want to discover the inherent groupings in the data, such as grouping customers by purchasing behavior.

Association

An association rule learning problem is where you want to discover rules that describe large parts of data, such as people that buy X also tend to buy Y.

Table 1 the Comparisons for types of Machine learning algorithms are tabulated as below.

Machine Learning	Data set	Data Type
Learning	Large	Integer, float
Unsupervised Learning	Large	Any
Active Learning	Low	Any
Reinforcement Learning	Medium	Any Type
Semi–Supervised Learning	Large	Any Type

Table 1 Machine learning comparison

Conclusion and Future work

Semantic web technologies offer a new approach to manage the collected information and processes by adding meaning to the web data. Ontology learning is one of the phases for managing information and develop the semantic construction in the semantic web.

To develop the semantic web system, it is essential to identify the need of the system, data needed in the system, the relationship of the data, and the concept of data and handling instances. Manual methods of knowledge management with semantic web system are time-consuming and expensive. Equations (1) & (2) are used as many times. By using Machine Learning methods with their corresponded algorithms, the data are manipulated by ontology construction. Based on data types and data quantity supervised learning is comparatively better than the other two.

Here, combining machine learning techniques and natural language processing techniques can be useful to build the Ontology in a semiautomatic way. In future automatically check the machine learning techniques with semantic.

References

1. Pradnya Gotmare," Methodology for Semi-automatic Ontology construction using Ontology Learning: A Survey," International Journal of Computer Applications (0975 – 8887) Emerging Trends in Computing 2016.
2. Patrick Hohenecker, Thomas Lukasiewicz," Deep Learning for Ontology Reasoning," arXiv: 1705.10342, Cornell University, May 2016.
3. Simon Bin, Patrick Westphal, Jens Lehmann, Axel Ngonga," Implementing Scalable Structured Machine Learning for Big Data in the SAKE Project,"
4. HamedHassanzadeh and Mohammed Reza Keyvanpour," A Machine Learning Based Analytical Framework for Semantic Annotation Requirement, "International Journal of web & Semantic Technology (IJWest) Vol.2, No.2, April 2011.
5. V.Devedzic, "Semantic Web and Education", Springer, 2006.
6. The Protégé Ontology Editor and Knowledge Acquisition System. [Online], Available: http://protege .stanford.edu/
7. Ian Horrocks, "DAML+OIL: A Description Logic for the semantic web", Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, 2002.

8. R. Sathya, Annamma Abraham," Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification," (IJARAI) International Journal of Advanced Research in Artificial Intelligence, Vol. 2, No. 2, 2013
9. Marta Garnelo, Kai Arulkumaran, Murray Shanahan,"Towards Deep Symbolic Reinforcement Learning", arXiv: 1609.05518v2 [cs.AI] 1 Oct 2016.
10. Euthymios Drymonas , Kalliopi Zervanou, and Euripides G.M. Petrakis," Unsupervised Ontology Acquisition from Plain Texts: the Onto Gain System,"
11. Guo Qingju Ji Wentian, Zhong Sheng, Zhou En, "The Analysis of the Ontology-based K-Means Clustering Algorithm", Proceedings of the 2nd International Conference on Computer Science and Electronics Engineering (ICCSEE 2013).
12. Linli Zhu, Yu Pan, Mohammad, Reza Farahani And Wei Gao, "Graph Laplacian Based Ontology Regularization Distance Framework for Ontology Similarity Measuring and Ontology Mapping",

A REVIEW ON BIG DATA ANALYTICS WITH HADOOP TECHNOLOGY

M.Saranya, MCA., M.Phil.,
Assistant Professor, Department of Information Technology
Madurai Sivakasi Nadars Pioneer Meenakshi Women's College, Poovanthi

Abstract
Big data is increasingly becoming a factor in production, marketplace competitiveness and the augment. Big data refers to enormous, varied, and often shapeless digital satisfied that is difficult to process using straight data administration tools and techniques. The term encompasses the complexity and diversity of data and data types, real-time data collection and dispensation needs, and the value that can be obtain by smart analytics. Huge challenges must be conquer if the payback is to be leveraged efficiently. Matter of concern besides increasing volumes of data, unreliable data structure and real-time dispensation include data safety, data privacy policy that are in vital need of reform and the increasing quality prospect of the stakeholders. There is an extensive lack of appropriate strategies to react to the digital uprising.
Keywords: *Big Data existence, convergence, competitiveness.*

Introduction

Data is as a compilation of large dataset that cannot be process using traditional computing techniques. Big Data is not just a data rather it has become a absolute subject which occupy various tools, technique and structure. Big Data is a term that refers to dataset whose volume (size), complexity and rate of growth (velocity) make them too difficult to captured, managed, processed or analyzed through conformist technology and tools such as relational databases. The term "Big Data" has recently been practical to datasets that grow so large that they become embarrassed to work with using conventional database running systems. They are data sets whose size is beyond the ability of normally used software tools and storage space systems to capture, store, direct, as well as process the data within a tolerable elapsed time. Big data sizes are always rising, currently range from a few dozen terabytes (TB) to many petabytes (PB) of data in a single data set. as a result, some of the difficulty linked to big data include capture, storage, search, sharing, analytics, and visualize. Hence, big data analytics is where advanced analytic technique is practical on big data sets. Analytics based on large data samples reveals and leverages business change. However, the larger the set of data, the more difficult it become to manage. Naturally, business benefit can normally be derived from analyze larger and more complex data sets that require real time or near-real time capability; however, this leads to a need for new data architectures, logical methods, and tools.

The Challenges of Big Data

Volume

Volume refers to amount of data. Volume stands for the size of the information how the data is large. The size of the data is representing in terabytes and petabytes.

Variety

Variety makes the data too big. The files comes in a variety of formats and of any type, it may be structured or formless such as text, audio, videos, log files and more.

Velocity

Velocity refers to the speed of data processing. The data comes at high speed. From time to time 1 minute is too late so big data is time sensitive.

Value

The potential value of big data is huge. Value is main source for big data because it is important for business, IT infrastructure system to store large amount of values in database.

Veracity

Veracity refers to noise, biases and irregularity. When we commerce with high volume, velocity and variety of data, the all of data are not going 100% accurate, there will be dirty data.

Hadoop: Solution for Big Data Processing

Hadoop is an Apache open source structure written in Java that allows distributed processing of large dataset across cluster of computers using straightforward programming model. Hadoop creates cluster of machines and organize work among them. It is intended to scale up from single servers to thousands of machines,

Each offering local calculation and storage. Hadoop consists of two component Hadoop Distributed File System (HDFS) [1] and Map Reduce Framework.

After the big data storage, comes the analytic dispensation. There are four critical supplies for big data processing. The first condition is fast data load. Since the disk and system traffic interferes with the query execution during data loading, it is essential to decrease the data load time. The second condition is fast query processing. In order to satisfy the necessities of heavy workloads and real-time needs, many queries are response-time critical. Thus, the data assignment structure must be able of retain high query dispensation speeds as the amounts of query rapidly increase. Additionally, the third must for big data processing is the highly efficient utilization of storage space. Finally, the fourth requirement is the strong adaptivity to highly dynamic workload patterns.

As big data sets are analyzed by different applications and users, for different purposes, and in various ways, the fundamental system should be extremely adaptive to unforeseen dynamics in data dispensation, and not specific to certain workload patterns. Map Reduce is a parallel encoding model, inspired by the “Map” and “Reduce” of useful languages, which is appropriate for big data dispensation. It is the core of Hadoop, and perform the data indulgence and analytics function. The basic idea of Map Reduce[4] is contravention a task down into stage and executing the stages in parallel in order to reduce the time needed to complete the task. The first phase of the Map Reduce job is to map input values to a set of key/value pairs as output. The “Map” function accordingly partitions large computational tasks into smaller tasks, and assigns them to the appropriate key/value pairs.

Thus, unstructured data, such as text, can be mapped to a structured key/value pair, where, for example, the key could be the word in the text and the value is the number of occurrences of the word. This output is then the input to the “Reduce” function. Reduce then performs the collection and combination of this output, by combine all values which share the same key value, to provide the final result of the computational task. The Map Reduce function within Hadoop depends on two different nodes: the Job Tracker and the Task Tracker nodes. The Job Tracker nodes are the ones which are responsible for distributing the mapped and reducer functions to the available Task Trackers, as well as monitoring the results. The Map Reduce job starts by the Job- Tracker transmission a portion of an input file on the HDFS to a map task, running on a node. On the other hand, the Task follower nodes actually

run the jobs and communicate results back to the Job Tracker. That communication between nodes is often from side to side files and directories in HDFS[4], so inter-node communication is minimized.

Figure 1 shows how the Map Reduce nodes and the HDFS work together. At step 1, there is a very large dataset counting log files, antenna data, or anything of the sorts. The HDFS stores replicas of the data represent by the blue, yellow, beige, and pink icons, across the data nodes. In step 2, the client defines and executes a map job and reduce job on an exacting dataset, and sends them both to the job Tracker. The job Tracker then distributes the jobs across the Task Trackers in step 3. The Task Tracker runs the mapped, and the mapped produces output that is then stored in the HDFS file system. Finally in step 4, the reduce job runs across the mapped data in order to construct the result.

Incentive: Our Data-Driven World

Advances in digital sensors, infrastructure, computation, and storage have created huge collection of data, capturing in order of value to business, science, administration, and society. For example, search engine companies such as Google, Yahoo!, and Microsoft have created an completely new business by capturing the information freely available on the World Wide net and as long as it to people in useful ways[2]. These companies collect trillions of bytes of data every day and repeatedly add new services such as satellite images, driving instructions, and image recovery. The societal payback of these services is immeasurable, having distorted how people find and make use of information on a daily basis. Some examples include:

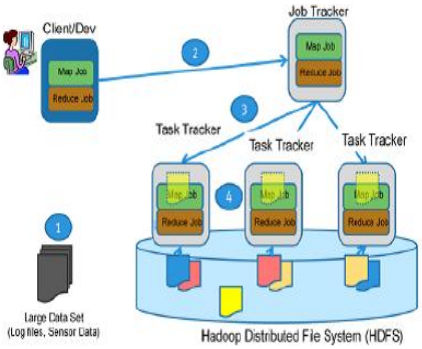


Fig. 1. MapReduce and HDFS

- Wal-Mart recently thin with Hewlett Packard to construct a data warehouse competent of storing 4 *petabytes* (4000 trillion bytes) of data, representing every single purchase recorded by their point-of-sale terminal (around 267 million transactions per day) at their 6000 stores worldwide. By applying *machine learning*[8] to this data, they can detect patterns indicating the effectiveness of their pricing strategies and advertising campaigns, and better manage their inventory and supply chains.
- Understanding the environment requires collecting and analyzing data from thousands of sensors monitoring air and water quality and meteorological situation, another example of science. This capacity can then be used to guide simulations of climate and groundwater models to create reliable methods to predict the effects of long-term trends, such as increased CO_2 emissions and the use of chemical fertilizers.
- Our cleverness agencies are being overwhelmed by the vast amount of data being composed through satellite images, signal intercept, and even from publicly available sources such as the Internet and news media. Finding and evaluate possible threats from this data require "connecting the dots" between manifold sources[6], e.g., to routinely match the voice in an intercept cell phone call with one in a video posted on a radical website.
- The compilation of all papers on the World Wide Web [6](several hundred trillion bytes of text) is prove to be a quantity that can be mined and process in many poles apart ways. For example, language conversions programs can be guided by statistical language models generate by analyze billions of papers in the source and target languages, as well as multilingual papers, such as the minutes of the United Nations. particular web crawlers scan for documents at different reading

levels to aid English-language education for first graders to adults. A conceptual network of noun-verb associations has been constructed based on word combinations found in web documents to guide a research project at Carnegie Mellon institution of higher education in which fMRI [5] images are used to detect how human brains store in order.

These are but a small sample of the ways that all facets of commerce, science, society, and national safety are being distorted by the ease of use of large amount of data and the means to take out new forms of considerate from this data.

Big-Data Technology

Sense, Collect, Store, and Analyze

The rising significance of big-data compute stems from advance in much different technology:

Sensors

Digital data are being generate by many different sources, counting digital imagers (telescopes, video cameras, MRI machines), chemical and biological sensors (microarrays, environmental monitors), and even the millions of persons and organizations generating web pages.

Computer Networks

Data from the many diverse sources can be together into enormous data sets via contained sensor networks, as well as the Internet [7].

Data Storage

Advances in attractive disk technology have radically decreased the cost of storing data. For example, a one-terabyte disk drive, investment one trillion bytes of data, costs around $100. As a position, it is predictable that if all of the text in all of the books in the Library of Congress could be transformed to digital form, it would add up to only around 20 terabytes.

Cluster Computer Systems

A new form of computer systems, consisting of thousands of "nodes," each having more than a few processors and disks, associated by high-speed local-area networks, have become the chosen hardware pattern for data-intensive computing systems. These clusters provide both the storage capacity for large data sets, and the compute power to systematize the data, to study it, and to respond to queries about the data from remote users. Compared with traditional high-performance computing[7] (e.g., supercomputers), where the focus is on maximize the raw compute power of a system, cluster computer are designed to maximize the reliability and competence with which they can direct and analyze very large data sets

Cloud Computing Facilities

The rise of large data centers and cluster computer has shaped a new business model, where businesses and persons can *rent* storage space and computing capacity, rather than making the large assets savings needed to construct and stipulation large-scale computer installation. For example, Amazon Web Services (AWS) provides both network-accessible storage priced by the gigabyte-month and compute cycles priced by the CPU-hour. Just as few organizations operate their own power plants, we can foresee an era where data storage and compute become utilities that are all over the place available.

Data Analytics Service Models

The Sass model offers total big data analytics applications to end users, who can exploit the cloud's scalability in both data storage and processing power to execute analysis on large or complex datasets. The PaaS model provides data analytics encoding suites and environments in which data mining developers can design scalable analytics services and application. Researchers can exploit the IaaS model to create a set of virtualized hardware and software capital for organization data analysis frameworks or applications. Developers can implement big data analytics services within each of these three models:

- **Data analytics software as a service**— provides a well-defined data removal algorithm or ready to-use knowledge discovery tool as an Internet service to end users, who can access it directly from side to side a Web browser;
- **Data analytics platform as a service**— provides a behind platform that developers can use to build their own data analytics application or extend existing ones without concern about the underlying communications or distributed computing Issues;
- **Data analytics infrastructure as a service**—provides a set of virtualized resources that developers can use as a computing infrastructure to run data mining applications or to put into practice data analytics system from scrape.

Conclusion

Big data is the new and that follow a series of rational stages in the development of the internet, such as individualization, the transfer of data to the cloud and the rapidly rising demand for digital mobility. It bridges the gap to what has evolved before. In principle, the idea is to combine different volumes of data with new data sets and to identify any patterns in this aggregated data using intelligent software, with the ultimate aim of drawing the right or possible conclusion from the findings. Once they have been compiling, primary data sets can be analyze any number of times for different purposes and for diverse stakeholders. The data function as a driver of innovation, creativity and out-of-the-box thoughts, and in an ideal world results in new business ideas, crop or services.

References

1. Andrew Pavlo, "A Comparison of Approaches to Large-Scale Data Analysis", SIGMOD, 2009. .
2. Apache Hadoop: http://Hadoop.apache.org
3. Dean, J. and Ghemawat, S., "Map Reduce: a flexible data processing tool", ACM 2010.
4. DeWitt & Stonebraker, "Map Reduce: A major step backwards", 2008.
5. Hadoop Distributed File System, http://hadoop.apache.org/hdfs
6. Hadoop Tutorial: http://developer.yahoo.com/hadoop/tutorial/module1.html
7. Big Data Analytics: Nada Elgendy, Ahmed Elragal
8. Scalable system for processing Big Data, Dawei Jiang, Gang Chen, Sai Wu

A STUDY ON REAL TIME DATA WAREHOUSE IMPLEMENTATION

K.Mahalakshmi, MCA., M.Phil.,
Head, Department of Information Technology
Madurai Sivakasi Nadars Pioneer Meenakshi Women's College, Poovanthi

Abstract
In this paper, I propose the concept of real-time data warehousing and its growth in recent years as organizations demand access to complicated part of data in real-time to produce analytics and make business decisions to gain more benefits. Real-time data warehousing systems differ significantly from traditional data warehousing systems, thus, presenting a unique part of organizational and operational challenges. The foundation for the research was to investigate whether adequate information is available regarding the organizational and operational difficulties of real-time data warehousing and whether that information is available to the database community. This expedition was done by gathering primary research, conducting a case study research design, and comparing, analyzing, and concluding facts from the two different categories of research.
Keywords: *Data Warehouse, Data Mart, Historical data storage.*

Introduction

The data warehouse is a method of storing historical and integrated data for use in decision support systems DSS. The data warehouse provides a source of integrated enterprise-wide historical data.[3]. Once the people have the data from the data warehouse, they can work with the data in order to make better decisions for their profit. Data stored in a data warehouse is available for messaged by users in which users can work with data in Excel, Power Pivot, pivot tables based off OLAP, cubes and Key Performance Indicators (KPIs). With the help of warehouse data, just about anyone can create complex models, build charts and calculations, manage a variety of reporting functions, analyze and make decisions.

The reason for developing an Enterprise Data Warehouse is that a reasonable data model exists deployed for users so that consistency is the key to consuming the data across the business. Regardless of the source of data, the data model gives users an accepted way to get the data they need to satisfy their purposes. This data model exists for use across the business to enhance the need for getting data in reports to make better decisions. Now more than ever, businesses rely on accurate, secure and up-to-date information in their reports to help their business operations and comply with increasingly stringent regulations. Having a prevalent source of data for users and taking away the burden reporting places on a transactional database can improve the efficiency and data sharing across the business. The solution of the data warehouse replaces Excel and other reporting platforms with a modern-day, centralized reporting and analysis solution. The ability to consolidate all significant and related data for a single version of the truth is fundamental to reporting accuracy.[2]

Data Marts

The final part of the dimensional model of data that comprises a data warehouse is the term known as a "data mart," which is the further classification of an existing star schema that is part of a data warehouse. The data mart is a subset of data that usually pertains to a part of the business and may need only be concerned by with a few fact tables and star joined dimension tables for queries and reports. Furthermore, once the data marts exist within the context of the data warehouse, users can

provide input for updates or further customization without affecting other data marts. There are many ways to describe the data mart, but one of the best definitions that I have come across is in terms that most anyone outside of an IT environment would be familiar. Since most people are familiar with a shopping mall, everybody was accustomed by with the fact that a shopping mall contains many stores, both large and small. Similarly, one can accurately describe that a data warehouse is similar to a shopping mall and a data mart is similar to a store within the mall and a store or collection of stores would represent star schema(s). Nonetheless, the coining and use of data marts, while simple and easy to understand to users and undoubtedly useful, have drifted away in recent years in favor of having the same structure but with less usage of this term. Regardless of whether one uses the term "data mart" or not, there is still going to be certain star schemas used by only certain groups, leaving other parts of the data warehouse untouched by them. Before getting too deep into the makeup of data marts, it is intriguing to understand that Kimball says, "When we first started using the data mart terminology in the 1990s, we were describing process-centric detailed databases that represented a subset of enterprise's overall data architecture" (Kimball, Lifecycle, 248). This remains very true as he describes data marts as being just a small part of the overall solution in a data warehouse. However, those in IT stopped using the term because the term data mart, in their eyes and in general was "hijacked to refer to independent, non-architected, department databases" (Kimball, 248). Such an instance refers to someone spinning up a separate database or small data warehouse just for his reporting needs alone and not keeping his eye on the bigger picture. Regardless of Kimball disowning the term data marts, their idea is still relevant, especially in the case of a large Enterprise Data Warehouse where users are going to need to find their data amongst a sea of star schemas.[7]

In order to create data marts, the very important prerequisite is the existence of the data warehouse, which is comprised of multiple star schemas, developed and named based off the dimensional model. It is true that it might just be easier for a group of users or a department perhaps to have its peculiar data warehouse for their resident transaction database and they could accomplish the same basic results with a data mart or small collection of data marts. The advantage of having the data warehouse or in this case, the Enterprise Data Warehouse, is that there may need to traverse across data marts into other data marts, which would not be possible in their own silo-type of data warehouse.[8] However, the largest advantage with an Enterprise Data Warehouse is that "you can shortcut the process of hunting the various candidate data sources, digging into their respective data environments to see what is there and what the quality looks like". Data marts offer a great advantage of having better control of data that users will need to find on a regular basis. Once the data marts exist as part of the large data warehouse solution, each area now has the responsibility of consuming the data within the mart for their own needs. We were reminded by "The most notable difference of a data mart from a data warehouse is that the data mart is created based on a very specific and predefined purpose and need for a grouping of certain data. A data mart is configured such that it makes access to relevant information in a specific area very easy and fast" (What Is a Data Mart, 1). If a basic user needs to go out to find their relevant star schema, it is very possible that within the scope of an Enterprise Data Warehouse, users could get lost or stumble across data that would not be relevant or applicable to their reporting needs. If the data mart presents the data properly for their use, they will quickly be able to develop the reports that they need to reap the benefits of a data warehouse in the first place.[11].

Literature Review

Bill Inmon defines data warehousing as a subject-oriented, integrated, time-variant, and non-volatile collection of data in support of management's decision-making process, in his book, Building the Data Warehouse.

In his definition, the data is: Subject-oriented as the warehouse is organized around the major subjects of the enterprise rather than the major application areas. This is reflected in the need to store decision-support data rather than application-oriented data. Integrated because of the coming together of source data from different enterprise-wide applications. The source data is often inconsistent using, for example, different formats. The integrated data source must be made consistent to present a unified view of the data to the users.[4] Time-variant because data in the warehouse is only accurate and valid at some point in time or over some time interval. The time-variance of the data warehouse is also shown in the extended time the data will help, the implicit or explicit association of time with all data, and the fact that the data represents a series of snapshots. Non-volatile as the data is not updated in real-time but is refreshed from operational systems on a regular basis. New data is always added as a supplement to the database, rather than a replacement. The database continually absorbs this new data, incrementally integrating it with the previous data.[1] According to Peter Robb and Carlos Coronel, authors of the textbook, ***Database Systems:*** *Design, Implementation, and Management,* the data warehouse is a read-only database optimized for data analysis and query processing.

Changing Data Sources in Data Warehouses

Extensive research has been performed by how data warehouses must be maintained by us in recent years. While data warehouses are designed to handle data changes in sources and the processing of this, often very little attention is allotted to the changes in the structure of data sources. The structure (tables, columns, data types, foreign key relations, etc.) of a database is described and defined in a so-called database schema. As Chen et al. state in [6], most of the data warehouse research assumes a static data warehouse schema, which is really not a valid assumption in an evolving environment. Sen and Sinha write in [5]. About the fact how data warehouse solutions often eventually fail because they are too complex and too expensive to change them to fit the evolving needs of the business. These evolving needs include aspects such as end-user improvements, data warehouse schema changes and other factors. One other factor that could be identified of evolving needs are that of changing source systems.

Chen et al. discuss the topic of view maintenance regarding data integration systems using the global-as-view approach. Various mapping techniques have been developed to specify how data of one schema is transformed to the other. One of these techniques is that of a global-as view query. This means that there is one global schema that defines how source systems (schemata) are mapped to gather the required information. Another technique is that of local-as view queries, where for each source system it is defined as a view over the global schema.

Distributed Data Warehouse Design

The design of a distributed database is aimed at the division of a database into several logic units known as fragments, which can be allocated to different sites. These fragments can be additionally replicated through the storage of redundant copies in two or more sites.

Aggregation

Data warehouses are structured according to different levels of aggregation. The lower level contains primitive data collected directly from the operational environment. On the other hand, the

higher level of the aggregation hierarchy stores highly summarized data. Between the lower and higher levels there may exist several intermediate levels, representing increasing aggregation degrees. Data stored in a level *n* correspond to some form of aggregation from data stored in a level *n-1*, and the lowest intermediate level utilizes data from the lower level as the basis for its aggregation.[12]

Quality in Data Warehouse

Data quality tools are used in data warehousing to ready the data and ensure that clean data populates the warehouses, thus enhancing the usability of the data warehouse. [9]

Data quality tools fall into one of the three categories: auditing, cleansing, and migration. Data auditing tools generally compare the data in the source database to set of business rules.

Data cleansing tools are used in intermediate staging area. A data cleansing tool cleans names, addresses and other data that can be compared to an independent source.

Data parsing breaks the record into atomic units that will be used by subsequent steps. Parsing includes placing elements of a record into the correct fields.

The data migration tool is used in extracting data from a source database and migrating data from staging area in the data warehouse.

Conclusion

In this paper, I explore the features that the data warehousing technology emerged as an alternative solution to traditional transaction processing commonly supported by conventional database systems. This technology is nowadays widely used by many companies to support the needs of management in the decision-making process. Algorithms and tools of a data warehousing environment allow selected data from heterogeneous and distributed information sources to be stored in a single data repository, known as a data warehouse. Since such a repository is specially organized to hold integrated, subject-oriented, historical, detailed and aggregated data, it is suitable to answer multidimensional queries and analyses from end users without accessing the original sources. Data warehousing provides the basis for corporate information environments. It guarantees efficiency and flexibility in the recovery of strategic information and maintains data with high quality and confidence.

References

1. Alec Pawling, “Data Warehousing For Social Networking and Human Mobility Research and Detecting Real-World Events In Space And Time Using Feature Clustering.”
2. Edward M. Leonard “Design and Implementation of an Enterprise Data Warehouse.”
3. Dale Hargens “Evaluation Real-Time Data Warehousing Challenges from a Theoretical and Practical Perspective.”
4. J. Kok “The design of a Delta Impact Analysis model for Data Warehouses.”
5. Sen, A.P. Sinha; A Comparison of data warehousing methodologies; Communications of the ACM, March 2005, Vol. 38, No. 3; p79-84.
6. S. Chen, X. Zhang, E.A. Rundensteiner; “A compensation-based approach for View Maintenance in Distributed Environments; IEEE Transactions on knowledge and data engineering.” Vol. 18, No. 8, August 2006; p1068-1081.
7. D. Theodoratos, T. Sellis “Data warehouse Configuration.”
8. W. H. Inmon, Richard D. Hackathorn “Using the Data Warehouse.”
9. M.Pamela Neely “Data Quality Tools for Data Warehousing- A small sample survey.”
10. Deodotta Boite “A Traffic Data warehousing and visualization scheme.”
11. Witold Abramowicz and Pawel Jan Kalczynski “Building and Taking Advantages of the Digital Library.”
12. Fernando Da Fonseca De Souza “Distributing Data Warehouse.”

AN OVERVIEW OF GOOGLE SERVICES ON CLOUD

M.Lydia Packiam Mettilda, MCA., M.Phil.,
Assistant Professor, Department of Information Technology
Madurai Sivakasi Nadars Pioneer Meenakshi Women's College, Poovanthi

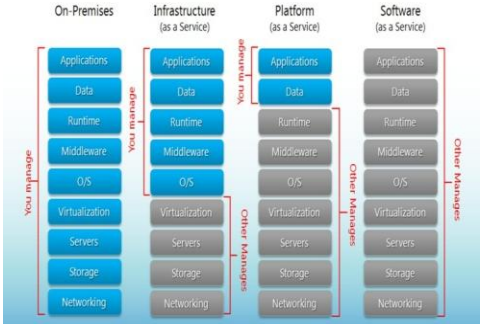
Abstract
In the current scenario, big and small organization around the world are adopting the Cloud Computing Technology, albeit in various models (public, private, hybrid), and this trend seems to be only increasing day by day. In today's digital world, Software-as-a-Service (SaaS) refers to a new and alternative way of accessing software, as opposed to more traditional methods of access. Google Cloud Platform was built to elevate people—to help people take their work further with Google technology and capabilities. Today faculty, researchers, administrators, and makers are taking advantage of Google Cloud Platform to build what's next. In this paper, analyzing an overview of various services of Google, particularly in the dominant model through SaaS.
Keywords: *Cloud, GCP, SaaS*

Introduction

Cloud Computing[1], often referred to as "the cloud," is the delivery of on-demand computing resources — everything from applications to data centers — over the internet on a pay-for-use basis. Cloud-based applications — or Software-as-a- Service — run on remote computers "in the cloud" that is owned and operated by others and that connect to users' computers via the internet and, usually, a web browser. Platform-as-a-Service provides a cloud-based environment with everything required to support the complete lifecycle of building and delivering web-based (cloud) applications — without the cost and complexity of buying and managing the underlying hardware, software, provisioning, and hosting. Infrastructure-as-a-Service provides companies with computing resources including servers, networking, storage, and data center space on a pay-per-use basis.

Google Cloud Platform

Google Cloud Platform[2] is a suite of public cloud computing services offered by Google. The platform includes a range of hosted services for compute, storage and application development that run on Google hardware. Google Cloud Platform services was accessed by software developers, cloud administrators and another enterprise IT professionals over the public internet or through a dedicated network connection.



Google Cloud Platform offers services for computing, storage, networking, big data, machine learning and the internet of things (IoT), as well as cloud management, security and developer tools. The core cloud computing products in Google Cloud Platform include:

- Google Compute Engine, which is an infrastructure-as-a-service (IaaS) offering that provides users with virtual machine instances for workload hosting.

- Google App Engine, which is a platform-as-a-service (PaaS) offering that gives software developers access to Google's scalable hosting. Developers can also use a software developer kit (SDK) to develop software products that run on App Engine.
- Google Cloud Storage, which is a cloud storage platform designed to store large, unstructured data sets. Google also offers database storage options, including Cloud Datastore for NoSQL non-relational storage, Cloud SQL for MySQL fully relational storage and Google's native Cloud BigTable database.
- Google Container Engine, which is a management and orchestration system for Docker containers that runs within Google's public cloud. Google Container Engine was based on the Google Kubernetes container orchestration engine.
- Google Cloud Platform offers application development and integration services. For example, Google Cloud Pub/Sub is a managed and real-time messaging service that allows messages was exchanged between applications.

Google continues to add higher-level services, such as those related to big data and machine learning, to its cloud platform. Google big data services include those for data processing and analytics, such as Google BigQuery for SQL-like queries made against multi-terabyte data sets. Also, Google Cloud Dataflow is a data processing service intended for analytics; extract, transform and load (ETL); and real-time computational projects. The platform also includes Google Cloud Dataproc, which offers Apache Spark and Hadoop services for Bigdata processing.

Examples of Google Services

Whereas in the past, the software would be purchased outright and loaded onto a device, SaaS refers to a subscription-based model where the software is hosted in the cloud and accessed via the internet. There are some benefits of this to consumers, whether that is individuals using software for private purposes, or businesses.

G-Suite

Google Suite (G Suite) [3] —formerly known as Google Apps—is Google's rebranded package of business productivity apps. It includes Gmail, Google Docs, Google Sheets, Google Slides, Google+, Google Calendar, Google Hangouts, and Google Forms. From creating decks for a marketing presentation on Google Slides to scheduling a monthly company-wide meeting on Google Calendar, using Google Suite is a no-brainer.

G Suite makes working together a whole lot easier. Make decisions faster, face to face. Use shared calendars to see when others are available and schedule meetings with automatic email invites. Collaborate in real-time. Store and share files in the cloud. Secure the data and devices.

A good example of SaaS is Google Docs. It is a productivity suite that is free for anyone to use. What all have to do is log-in and instantly have access to a word processor, spreadsheet application, and presentation creator. Google's online services are managed directly from the web browser and require zero installation. Anyone can access Google Docs from any computer or mobile device with a web browser.

Google App Engine

Google App Engine is a Platform as a Service (PaaS) product that provides Web app developers and enterprises with access to Google's scalable hosting and tier 1 Internet service. The App Engine requires

that apps be written in Java or Python, store data in Google BigTable and use the Google query language. Google App Engine is free up to a certain amount of resource usage. Users exceeding the per-day or per-minute usage rates for CPU resources, storage, number of API calls or requests and concurrent requests can pay for more of these resources.

Google Compute Engine

Google Compute Engine (GCE) is an Infrastructure as a Service (IaaS) offering that allows clients to run workloads on Google's physical hardware. Google Compute Engine provides a scalable number of virtual machines (VMs) to serve as compute clusters for that purpose. GCE was managed by a RESTful API, command line interface (CLI) or Web console. Compute Engine is a pay-per-usage service with a 10-minute minimum.

Conclusion

In this paper, presented an overview of various Google services in the part of Cloud. During this, provides a good, quick resource about the state of the Google Cloud Platform and everything that Google has to offer with their cloud solution. Ultimately, only by using and testing the platform will gain a complete understanding and will meet off-premises computing goals.

References

1. https://www.ibm.com/cloud/learn/what-is-cloud-computing
2. http://searchcloudcomputing.techtarget.com/definition/Google-Cloud-Platform
3. https://www.workato.com/blog/2017/06/5-useful-google-suite-features/#.WmvJJ
4. http://searchaws.techtarget.com/definition/Google-Compute-Engine
5. http://www.hostingadvice.com/how-to/iaas-vs-paas-vs-saas/
6. https://cloud.google.com/appengine/
7. https://console.cloud.google.com/appengine/create?lang=java&project=modular-robot-194016
8. https://www.interoute.com/what-saas

ISSUES OF CYBERCRIME SECURITY

B.Shameera, III B.Sc., (Computer Science)
Madurai Sivakasi Nadars Pioneer Meenakshi Women's College

Abstract
The purpose of understanding the Cybercrime is a remarkable thing, Challenges and Legal Response is to assist everyone in understanding the legal aspects of cyber security and to help make consistent or compatible legal essential supporting structure. Also discusses the issue of cyber crime, including the types, methods and effects of Criminal activities carried out on a computers, internet or network. With the high speed of technological developments, our life is becoming more modern and digitized. Be it business, education, shopping or banking transactions everything is in the Cyberspace. There are some things likely to cause damage or danger posed by this incredible rise in digitization which is creating a new set of global relate to called as Cybercrime
Keywords *Security, Network Security, Computer Privacy, Cybercrimes.*

Introduction

Cyber Crimes concern the word cyber which cause to take place the computer network, using which one can perform any activity in the real-time world. Cyber Crimes such as committing fraud, illegal trade in child pornography and a person possessing a highly developed intellect property, stealing a close similarity or affinity, or violating privacy use digital data from Computer systems and other electronic devices. The arrival of notable thing of computers and the expansion of the Internet made likely the thing that has been achieved successfully of the wide range of improvement in research, surgery, expertise, and communication. Unfortunately, computers and the Internet have furthermore supplied a new straightforward environment for crime. Cybercrime is roughly characterized as distinctive nature of a wicked or illegal act through the use of a computer or the Internet. The Internet has been described as the "collectively the myriad of computer and telecommunications amenities, encompassing gear and functioning programs. The Internet is a considerable or relatively great size or capacity computer material made of a network of wire or a string of connections of computers that are fastened or connected.

Purpose: The purpose of the study is to determine the impact of cybercrimes on network security and to firmly decide at what level network security can make worse or less desirable cyber-crimes.

Impacts of Cybercrime

The main is to firmly decide the marked effect or influence of cybercrime on networks and to come to an end how network security reduces the Cyber-crimes. There are a lot of cases of Computer Assisted crime where a computer is the tool or device for committing the illegal activities. Some of them have been discussed below:

1. Data Piracy: This involves reproduction of digital data and achieved without great distribution of print, graphics, sound and multimedia combinations even the use of copyrighted material either for personal use.
2. Pornography/Child Pornography: It is the unethical and illegal distribution of sexually implicit material especially involving children.

3. Internet Time Thefts: Phishing, spoofing or spam (unsolicited mail) wherein a perpetrator sends fictitious report send by post which appears official causing the victim to release personal information.
4. Online Credit Card Fraud, E-Bank Theft: without official authorization of credit card number for online acquire by paying it or bank account details where a person who carries out a harmful, illegal or immoral act diverts funds to account accessible to criminal.
5. Forgery: It includes the reproduction of documents, certificates, identity thefts and fake currency.
6. Altering Websites: Here the hacker deletes some pages of a website, uploads new pages with the similar name and controls the messages conveyed by the website.
7. Cyber Terrorism: It involves E-murder or homicide or suicide or Spyware.

Causes of Cybercrime

1. Ease of Access: The problem encountered in guarding a computer system against unauthorized access is that there is every possibility of treating with irreverence or disrespect the technology by using codes, recorders, pins, retina image, etc without permission. that can be used to fool biometric systems and bypass firewalls to get past many a security system.
2. Cyber Hoaxes: Cyber Crimes can be performed just to cause danger or damage one's reputation. Cyber hoaxes are one of the most dangerous of all causes. They believe in fighting their cause and want their goal to be achieved. They are called cyber- terrorists.
3. Negligence: There are possibilities of not paying attention to protecting the system. This negligence gives the criminals control to damage the computer.
4. Revenge or Motivation: The intense and selfish desire to master the complicated system with a desire to inflict loss to the victim. Revenge or motivation includes youngsters or those who are driven by lust to make quick money and they make unauthorized alterations with data like e-commerce, e-banking or fraud in transactions.
5. Poor Law Enforcing Bodies: Due to lack of cyber laws of many countries, many criminals get away without being punished.
6. Cyber Crimes Committed for Publicity or Recognition: Committed by youngsters where they just want to be noticed without hurting someone's sentiments.

Four Levels of Cybercrime Victims

They are:

1. Gullible
2. Desperados
3. Inexperienced
4. Unlucky people

Convention on Cybercrime

Offense against the confidentiality, integrity, and availability of computer data and systems;

1. computer-related offenses
2. content-related offenses
3. copyright-related offenses

Defending Against Cyber Crimes

1. To Prevent Against Exploitation: Periodic scanning for spyware, adware, and bots (software robots) shall be conducted with anti-spyware programs that detect these malicious programs. Provision of security awareness training to personnel on an annual basis that, in part, cautions against downloading software programs from the Internet without appropriate agency approval
2. Virus Protection: To minimize virus hoax Virus detection programs and practices shall be implemented throughout agencies All agencies shall be responsible for ensuring that they have current software on their network to prevent the introduction or spreading or promoting of computer viruses by using of antivirus software, committing backups more than a time at short intervals of time on data files, using write-protected program media, such as diskettes or CDROMs, checking the source of software before installing it all CDs or other media brought from home or any other outside.

Development of Computer Crime and Cybercrime in 21st Century

As in each preceding decade, new trends in computer crime and cybercrime continued to be discovered in the 21st century. The first decade of the new millennium was dominated by new, highly sophisticated methods of committing crimes, such as "phishing", and "botnet attacks", and the emerging use of technology that is more difficult for law enforcement to handle and investigate, such as it is not only the methods that changed but also have the strong effect . As offenders came able to automate attacks, the number of offences increased. Countries and regional and international organizations have responded to the growing challenges and given response to Cybercrimes high priority.

Terrorists Use Internet For

It is well known that terrorists use ICTs and the Internet for:

- Propaganda
- Information gathering
- Preparation of real-world attacks
- Publication of training material
- Communication
- Terrorist financing
- Attacks against critical infrastructures.

Crime Statistics

1. The US Internet Complaint Center reports a 22.3 percent increase in complaints submitted relating to cybercrime compared with 2008
2. German Crime Statistics indicate that the overall number of Internet-related crimes increased in 2009 by 23.6 per cent compared with 2008

Benefits of Network Security

1. Prevents unauthorized users from accessing your network.
2. Provides transparent access to Internet-enabled users.
3. Ensures that sensitive data is transferred safely by the public network.
4. Help your managers to find and fix security problems.
5. Provides a comprehensive system of warning alarms attempt to access your network.

Advantages and Risks

The growth of the information society is accompanied by new and serious threats.21 Essential services such as water and electricity supply now rely on ICTs. Cars, traffic control, elevators, air conditioning and telephones also depend on the smooth functioning of ICTs. Attacks against information infrastructure and Internet services now have the potential to harm society in new and critical ways. Attacks against information infrastructure and Internet services have already taken place.

The Use of Encryption Technology

Perpetrators are increasingly encrypting their messages. Law-enforcement agencies note that offenders are using encryption technology to protect information stored on their hard disks, seriously hindering criminal investigations.

Prevention

Apart from his own mentality and the strength of his motivations, the criminal also needs to see the path of crime ahead of him clear of a thing which stops or hinders. If every single individual were to put up obstacles of their, no matter how small, the crime path will seem to be far less lucrative in the eyes of an even most desperate criminal. The fight against cybercrime must start with preventing it in the first place.

Users

The individual should be proactive, not reactive. You do not have to remain at the receiving end of crime forever. The fight against cybercrime starts in your very own home. Person should not reply any e-mail from unknown persons, they should learn to report spam mails to the e-mail server or, any know cybercrime research sites. If there is one thing that makes performing cybercrime lucrative, it is the fact that victims rarely have the required knowledge or presence of mind to handle the situation.

Conclusion

In conclusion, it can be said that attacks on machines connected to the Internet have increased by 260% since 1994, with an estimated loss of 1,290 million dollars yearly in the U.S. It has been found educating the users on being safe online needs to be accepted generally which had been done for a long time to know about the pros and cons of the web before entering it. There is no doubt that the Internet offers criminals several opportunities. Information is the best form of protection. With advances in technology, no one is safe from an attack by "hackers. Currently, it is relatively easy to gain control of a machine on the Internet that has not been adequately protected. Companies invest a sufficiently great portion of their money in protecting their information since the loss of data which cannot be replaces when lost or damaged is a real threat to their business.

References

1. Dacey, Raymond & Gallant, Kenneth S. (1997) "Crime control and harassment of the innocent," Journal of Criminal Justice, Elsevier, vol. 25(4), pages 325- 334.
2. Panu Poutvaara & Mikael Priks (2005) "Violent Groups and Police Tactics: Should Tear Gas Make Crime Preventers Cry?, " CESifo Working Paper Series 1639, CESifo Group Munich.
3. Ying-Chieh Chen, Patrick S. Chen, Jing-Jang Hwang, Larry Korba, Ronggong Song, George Yee, (2005) "An analysis of online gaming crime characteristics", Internet Research, Vol. 15 Iss: 3, pp.246 - 261
4. Whitman, Michael E. & Mattord, Herbert J. Principles of information security (2nd ed.)

5. Aghatise E. J. (2006): Level of Awareness of Internet Intermediaries Liability. (HND Project work) Unpublished. Auchi Polytechnic, Auchi, Edo State, Nigeria.
6. Longe, O.B. (2004): Proprietary Software Protection and Copyright issues in contemporary Information Technology. (M.Sc Thesis) Unpublished. Federal the University of Technology, Akure, Nigeria.
7. Smith, R. G., Holmes, M. N. &Kaufmann, P. (1999): Nigerian Advance Fee Fraud., Trends and Issues in Crime and Criminal Justice, No. 121, Australian Institute of Criminology, Canberra (republished in The Reformer February 2000, pp. 17-19).
8. Sylvester, Linn (2001): The Importance of Victimology in Criminal Profiling. Available online at: http://isuisse.ifrance.com/emmaf/base/impvic.html

INFORMATION STORAGE WITH CLOUD SERVICE

M.Gayathri, III B.Sc., (Computer Science)
Madurai Sivakasi Nadars Pioneer Meenakshi Women's College

Abstract
Cloud computing is the development of parallel computing, distributed computing, grid computing and virtualization technologies which define the shape of a new era. Cloud computing is an emerging model of business computing. Security and privacy issues present a strong barrier for users to adapt into cloud computing systems. Cloud is not a particular product, but a way of delivering IT services that are consumable on demand, elastic to scale up and down as needed, and follow a pay-for-usage model. Security is not robust and consistent, that flexibility and the advantages of cloud computing has to offer will have little credibility.
Keywords: C*loud service, Architecture, IaaS, SaaS, PaaS, Cloud security.*

Introduction

Cloud computing is a complete new technology. It is the development of parallel computing, distributed computing grid computing, and is the combination and evolution of Virtualization, Utility computing, Software-as-a-Service (SaaS), Infrastructure-as-a-Service (IaaS) and Platform-as-a-Service (PaaS). Cloud is a metaphor to describe web as a space where computing has been pre installed and exist as a service of data, operating systems, applications, storage and processing power exist on the web ready to be shared. To users, cloud computing is a Pay-per-Use-On-Demand mode that can conveniently access shared IT resources through the Internet. Where the IT resources include network, server, storage, application, service and so on and they can be deployed with much quick and easy manner and least management and also interactions with service providers. Cloud enables the consumers of the technology to think of computing as effectively limitless, of minimal cost, and reliable, as well as not to be concerned about how it is constructed, how it works, who operates it, or where it is located.

Cloud Service

The term cloud computing is rather a concept which is a generalized meaning evolved from distributed and grid computing. The straightforward meaning of cloud computing refers to the features and scenarios where total computing could be done by using someone else's network where ownership of hardware and soft resources are of external parties.

The computing world has been introduced with a number of terminologies like SaaS (Software as a Service), PaaS (Platform as a Service) and IaaS (Infrastructure as a Service) with the evolution of cloud computing. In a pervasive meaning within the context of computer networks, infrastructure could be thought of as the hardware as well as their alignment where platform is the operating system which acts as the platform for the software. The concept of cloud based services is hierarchically built from bottom to top in the order of IaaS, PaaS and SaaS. This is merely the level of abstraction that defines the extent to which an end-user could 'borrow' the resources ranging from infrastructure to software – the core concern of security and the fashion of computing are not affected by this level of abstraction. Virtualization is an inevitable technology that is highly coupled with the concept of cloud computing. It is the virtualization technology that complements cloud services in the form of PaaS and SaaS where one physical infrastructure contains services or platforms to deliver a number of cloud users

simultaneously. This leads to the addition of total security aspects of virtualization technology on top of the existing security concerns and issues of cloud computing.

Architecture

Cloud service models are commonly divided into SaaS, PaaS, and IaaS that exhibited in the cloud infrastructure. It's helpful to add more structure to the service model stacks:

a. Software as a Service (SaaS): Cloud consumers release their applications in a hosting environment, which can be accessed through networks from various clients (e.g. Web browser, PDA, etc.) by application users. Cloud consumers do not have control over the cloud infrastructure that often employs multi-tenancy system architecture, namely, different cloud consumers applications are organized in a single logical environment in the SaaS cloud to achieve economies of scale and optimization in terms of speed, security, availability, disaster recovery and maintenance. Examples of SaaS include SalesForce.com, Google Mail, Google Docs, and so forth.
b. Platform as a Service (PaaS): PaaS is a development platform supporting the full "Software Lifecycle" which allows cloud consumers to develop cloud services and applications (e.g. SaaS) directly on the PaaS cloud. Hence, the difference between SaaS and PaaS is that SaaS only hosts completed cloud applications whereas PaaS offers a development platform that hosts both completed and in-progress cloud applications. This requires PaaS, in addition to supporting application hosting environment, to possess development infrastructure including programming environment, tools, configuration management, and so forth. An example of PaaS is Google AppEngine.
c. Infrastructure as a Service (IaaS): Cloud consumers directly use IT infrastructures (processing, storage, networks and other fundamental computing resources) provided in the IaaS cloud. Virtualization is extensively used in IaaS cloud in order to integrate/decompose physical resources in an ad-hoc manner to meet growing or shrinking resource demand from cloud consumers. The basic strategy of virtualization is to set up independent virtual machines (VM) that are isolated from both the underlying hardware and other VMs. Notice that this strategy is different from the multi-tenancy model, which aims to transform the application software architecture so that multiple instances (from multiple cloud consumers) can run on a single application (i.e. the same logic machine). An example of IaaS is Amazon's EC2.
d. Data as a Service (DaaS): The delivery of virtualized storage on demand becomes a separate Cloud service - data storage service. Notice that DaaS could be seen as a special type IaaS. The motivation is that on-premise enterprise database systems are often tied in a prohibitive upfront cost in dedicated server, software license, post-delivery services and in-house IT maintenance. DaaS allows consumers to pay for what they are actually using rather than the site license for the entire database. In addition to traditional storage interfaces such as RDBMS and file systems, some DaaS offerings provide table-style abstractions that are designed to scale out to store and retrieve a huge amount of data within a very compressed timeframe, often too large, too expensive or too slow for most commercial RDBMS to cope with. Examples of this kind of DaaS include Amazon S3, Google BigTable, and Apache HBase, etc.

Confidentiality in Cloud Computing

Confidentiality is one of the most important security mechanisms for users' data protection in the cloud. It includes encryption of the plaintext in cipher text before the data is stored in the cloud. This

technique protects the users' data and even cloud service providers cannot modify or read the content that is stored in this way in the cloud.

This kind of protection is offered from Dell data protection and encryption where users' data is protected when it is stored on the external drive or media. Encryption could be done either using software or hardware. Great benefit of this kind of protection is that users don't need to bother with the enforce policies of Dell data protection and encryption. Dell also uses Transparent File Encryption to control the users that are accessing the data.

Wuala cloud is another vendor that enables encryption for the data in the cloud. Encryption is enabled here before personal computers are sending the data to the cloud. This is excellent protection because even the provider cannot access the data.

Confidentiality is also provided by the vendor Online Tech which obtains confidentiality in the cloud computing using encryption methods (like Full Disk Encryption) that encrypt stored data on hard disk throughout the booting process. Whole Disk Encryption is also used for encrypting the data with the well known AES (Advanced Encryption Standard) algorithm. If the device that is using cloud computing technology is lost or stolen there is also a bit locker password which protects the data on the lost or stolen device.

Hence, we can conclude in this section that confidentiality is very important for protecting the data in the cloud and different vendors offer different security techniques for ensuring the confidentiality. Per example, DELL offers hardware and software based encryption, as well as transparent file encryption. The benefits of this kind of encryption techniques are that they are easy to implement and intervention of the user is not needed. Wuala is using encryption techniques on personal computers and this method for encryption in the cloud gives advantage of the users for accessing the data. Online Techoffers Full Disk Encryption and Whole Disk Encryption in order to enable confidentiality of the data in the cloud. Benefits of these encryption methods are that data that are partitioned could be decrypted and data is encrypted at rest.

Cloud Security

Cloud computing can provide infinite computing resources on demand due to its high scalability in nature, which eliminates the needs for Cloud service providers to plan far ahead on hardware provisioning. As cloud computing refers to both the applications delivered as services over the Internet and the infrastructures (i.e., the hardware and systems software in the data centers) that provide those services. Based on the investigation security and privacy concerns provided by companies nowadays are not adequate, and consequently result in a big obstacle for users to adapt into the cloud computing systems. Hence, more concerns on security issues, such as availability, confidentiality, data integrity, control, audit and so on, should be taken into account.

Cloud computing allows providers to develop, deploy and run applications that can easily grow in capacity (scalability), work rapidly (performance), and never (or at least rarely) fail (reliability), without any concerns on the properties and the locations of the underlying infrastructures.

Cloud computing systems can achieve the following five goals together:

1. Availability: The goal of availability for cloud computing systems (including applications and its infrastructures) is to ensure its users can use them at any time, at any place. As its web-native nature, cloud computing system enables its users to access the system (e.g., applications, services) from anywhere. This is true for all the cloud computing systems (e.g., DaaS, SaaS, PaaS, IaaS, and etc.). Required to be accessed at any time, the cloud computing system should be severing all the

time for all the users (say it is scalable for any number of users). Two strategies, say hardening and redundancy, are mainly used to enhance the availability of the cloud system or applications hosted on it.

2. Confidentiality: It means keeping users' data secret in the cloud systems. There are two basic approaches (i.e., physical isolation and cryptography) to achieve such confidentiality, which are extensively adopted by the cloud computing vendors.
3. Data integrity: In the cloud system means to preserve information integrity (i.e., not lost or modified by unauthorized users). As data are the base for providing cloud computing services, such as Data as a Service, Software as a Service, Platform as a Service, keeping data integrity is a fundamental task.
4. Control: In the cloud system means to regulate the use of the system, including the applications, its infrastructure and the data.
5. Audit: It means to watch what happened in the cloud system. Auditability could be added as an additional layer in the virtualized operation system (or virtualized application environment) hosted on the virtual machine to provide facilities watching what happened in the system. It is much more secure than that is built into the applications or into the software themselves, since it is able watch the entire access duration.

Future Developments

There are also some available surveys concerning security issues in Cloud computing namely the ones that can impair integrity, availability, and confidentiality. Using only firewall devices will not help solve these problems. Consequently, examine proposals that incorporate the joint use of IDS (Intrusion Detection Systems) and IPS (Intrusion Prevention Systems). On one hand, the proposals based on signature detection offer the advantage of minimal response time and human intervention but have the disadvantage of not being able to detect previously unknown ('zero day') attacks. On the other hand, anomaly detection proposals have opposite functional characteristics in comparison with signature-based ones. Hybrid cloud IDPS schemes should be investigated for use in future systems.

Future cloud systems should be able to detect and prioritize simultaneous attacks in terms of their negative impact on the system performance. Then, these systems need to put into action prioritized corrective measures to limit the destructiveness of the more dangerous attacks. In addition, the security solutions should scale or adjust network node numbers, the node heterogeneity (e.g. a federated cloud system), and traffic load, to offer a satisfactory service. It is also worth noting that there is a trade-off between performance and the level of security adopted. Clearly, higher security levels will necessitate more checking, and consequently there will be fewer resources for regular customer use. It is therefore advisable to apply the minimally appropriate set of policies by means of self-managing and self-learning.

Cloud users would also need to feel confident that their data privacy is guaranteed when they upload the data to the cloud.

A significant piece of research is currently being carried out in the European FP7 project SECCRIT (Secure Cloud Computing for Critical Infrastructure IT), which addresses technical and legal issues in the context of cloud security. This is a multidisciplinary research project with the mission to analyze and evaluate cloud computing technologies with respect to security risks in sensitive environments, and to develop methodologies, technologies, and best practices for creating secure, trustworthy, and high assurance cloud computing for critical infrastructure IT." Also, the project is investigating relevant European legal frameworks with the aim of establishing guidelines for using cloud services in the

critical infrastructure sector. Otherwise, the use of cloud in this sector, where stringent regulatory and legal requirements exist, will continue to be severely limited. Furthermore, clear guidelines are needed on how to deal with liability issues following any service failures.

Very recently a new cloud service model is winning a considerable importance, the Data as a Service (DaaS), which we discuss in the following subsection.

Security Risks of an Emerging Cloud Service Model: Data as a Service

The typical usage scenario of this model is the one where the user data is outsourced to the cloud (e.g. Dropbox). However, the data owners lose control over their data because the cloud provider becomes a third party service provider. An initial solution to ensure the data privacy is to encrypt it before exporting it to the cloud. A legacy solution to this issue is based on symmetric key encryption but it is not secure when a revoked user rejoins the system. In this way, proposes a homomorphic encryption and proxy re-encryption scheme that prevents leakage of data privacy when a revoked user rejoins the system. This solution also prevents the collusion between a revoked user and the cloud provider. It also supports secure query processing over the encrypted data already stored in a federation of clouds.

Conclusion

Cloud computing is becoming a hugely attractive paradigm, especially for large enterprises. Cloud service providers need to inform their customers on the level of security that they provide on their cloud. In this paper, we first discussed various models of cloud computing, security issues and research challenges in cloud computing. Data security is major issue for Cloud Computing. There are several other security challenges including security aspects of network and virtualization. This paper has highlighted all these issues of cloud computing. We believe that due to the complexity of the cloud, it will be difficult to achieve end-to-end security. New security techniques need to be developed and older security techniques needed to be radically tweaked to be able to work with the clouds architecture. Cloud Computing initiatives could affect the enterprises within two to three years as it has the potential to significantly change IT. Cloud computing has enormous prospects, but the security threats embedded in cloud computing approach are directly proportional to its offered advantages. The security issues for cloud computing are not related to the technical and direct security breach only; a number of social inconsistency might also be resulted even without any 'hard' security breach having taken place. The distributed and dispersive processing, transmission and storage features are behind reason. One such example is the obtaining of digital evidences. The evolution of cloud computing might significantly affect the collection and retention of digital evidence.

References

1. T. Dillon, C. Wu, and E. Chang, "Cloud Computing: Issues and Challenges," 2010 24th IEEE International Conference on Advanced Information Networking and Applications (AINA), pp. 27-33, DOI= 20-23 April 2010
2. M. Q. Zhou, R. Zhang, W. Xie, W. N. Qian, and A. Zhou, "Security and Privacy in Cloud Computing: A Survey," 2010 Sixth International Conference on Semantics, Knowledge and Grids(SKG), pp.105-112, DOI= 1-3 Nov. 2010
3. Bisong, A. and Rahman, S.S.M. (2011). An Overview of the Security Concerns in Enterprise Cloud Computing. International Journal of Network Security & Its Applications, 3(1), 30-45.

doi:10.5121/ijnsa.2011.3103

4. Atayero, A.A. and Feyisetan, O. (2011). Security Issues in Cloud Computing: The Potentials of Homomorphic Encryption. Journal of Emerging Trends in Computing and Information Sciences, 2(10), 546-552.
5. S. Subashini, V. Kavitha, "A survey on security issues in service delivery models of cloud computing"; Journal of Network and Computer Applications, Vol. 34(1), pp 1–11, Academic Press Ltd., UK, 2011, ISSN: 1084-8045.
6. V. Krishna Reddy, B. Thirumal Rao, Dr. L.S.S. Reddy, P.Sai Kiran "Research Issues in Cloud Computing " Global Journal of Computer Science and Technology, Volume 11, Issue 11, July 2011.
7. S. Zhang, S. F. Zhang, X. B. Chen, and X. Z. Huo, "Cloud Computing Research and Development Trend," In Proceedings of the 2010 Second International Conference on Future Networks (ICFN '10). IEEE Computer Society, Washington, DC, USA, pp. 93-97. DOI=10.1109/ICFN.2010. 58.
8. J. J. Peng, X. J. Zhang, Z. Lei, B. F. Zhang, W. Zhang, and Q. Li,"Comparison of Several Cloud Computing Platforms," 2009 SecondInternational Symposium on Information Science and Engineering(ISISE '09). IEEE Computer Society, Washington, DC, USA, pp. 23-27,DOI=10.1109/ISISE.2009.94.
9. M. M. Alabbadi, "Cloud Computing for Education and Learning: Education and Learning as a Service (ELaaS)," 2011 14th International Conference on Interactive Collaborative Learning (ICL), pp. 589 – 594, DOI=21-23 Sept. 2011.
10. Anthony T. Velte, Toby J. Velte, Robert Elsenpeter, Cloud Computing: A Practical Approach (McGraw Hill Publications, pp 135 – 144, 2010).
11. Daniel J. Abadi, Data Management in the Cloud: Limitations and Opportunities, IEEE Data Engineering Bulletin, Volume 32, March 2009, 3-12.
12. Linda Xu, Miklos Sandorfi and Tanya Loughlin, Cloud Storage for Dummies (Wiley Publishing, pp. 5-24, 2010).
13. Khorshed, T.M., Ali, A.B.M.S. and Wasimi, S.A. (2012). A survey on gaps, threat remediation challenges and some thoughts for proactive attack detection in cloud computing. Future Generation Computer Systems, 28, 833–851. doi:10.1016/j.future.2012.01.006
14. Sharma, S. And Mittal, U. (2013). Comparative Analysis of Various Authentication Techniques in Cloud Computing. International Journal of Innovative Research in Science, Engineering and Technology, 2(4), 994-998.
15. K. Vieira, A. Schulter, C. B. Westphall, and C. M. Westphall, "Intrusion detection techniques for Grid and Cloud Computing Environment," IT Professional, IEEE Computer Society, vol. 12, issue 4, pp. 38-43, 2010.

DATA MINING TECHNIQUES IN SOCIAL MEDIA

I.Razul Beevi, MCA.,
Assistant Professor of Information Technology
Madurai Sivakasi Nadars Pioneer Meenakshi Women's College

Abstract
Today, the use of social networks is growing ceaselessly and rapidly. More alarming is the fact that these networks have become a substantial pool of unstructured data that belong to a host of domains, including business, governments, and health. The increasing reliance on social networks calls for data mining techniques that are likely to facilitate reforming the unstructured data and place them within a systematic pattern. The goal of the present survey is to analyze the data mining techniques that were utilized by social media networks between 2003 and 2015. Espousing criterion-based research strategies, 66 articles were identified to constitute the source of the present paper. After a careful review of these articles, we found that 19 data mining techniques have been used with social media data to address 9 different research objectives in 6 different industrial and services domains. However, the data mining applications in the social media are still raw and require more effort by academia and industry to adequately perform the job. We suggest that more research was conducted by both the academia and the industry since the studies have done so far are not sufficiently exhaustive of data mining techniques.
Keywords: *Data Mining, Social Media, Social Media Networks Analysis, Survey*

Introduction

The Social network is a dedicated website enabling the user to communicate with each other by their post, videos, comments, etc. Also, they are web-based services that allow individuals are creating a public profile in a domain such that they can communicate with other users within that network. The Social network has improved on the concept of the technology of Web 2.0, by enabling the formation and exchange of User-Generated Content. The Social network is a graph which includes of nodes and links representing the social relations on social network websites [1]. A node includes many entities and the relationships between them forming links. Social networks are important sources of online interactions and contents sharing, subjectivity [2], approaches, evaluation, influences, observations, feelings, opinions and sentiments expressions bear Out in the text, reviews, blogs, discussions, news, remarks, reactions, or some other documents. Before the advent of the social network, the homepages was popularly used in the late 1990s which made it likely for average internet users to share information. However, the activities on the social network in recent times seem to have transformed the World Wide Web (www) into its intended innovative creation. Social network platforms enable rapid information exchange between users regardless of the location. Many organizations, individuals and even government of countries now pursue the activities on the social network. The network enables big organizations, celebrities, government official and government bodies to obtain knowledge on how their audience reacts to postings that concerns them out of the enormous data generated on the social network. The network permits the effective collection of large-scale data which gives climb to major computational challenges. Yet, the application of efficient data mining techniques has made it possible for users to discover valuable, accurate and three dominant disputes with useful social knowledge from social network data. Data mining techniques are capable of handling network data viz., size, noise and, dynamism. The enormous nature of social network datasets requires information processing to be done automatically for analyzing within a reasonable time. Amusingly data mining techniques require huge data sets to mine the remarkable patterns from data; social networking sites appear to be perfect to

mine with data mining tools. This form an enabling factor for advanced results for a search in searching engines and also helps in better understanding of social data for research and organizational functions [3]. Data mining tools used for a survey in this paper ranges from unsupervised, semi-supervised to supervised learning.

Social Network Background

During the last decade, social network have become not only popular but also affordable, and universally-acclaimed communication means that has thrived in making the world a global village. Social network sites are commonly known for information dissemination, personal activities posting, product reviews, online pictures sharing, professional profiling, advertisements and opinion/sentiment expression. News alerts, breaking news, political debates and government policy are also posted and analyzed on social network sites. It was observed by more people are becoming interested in and relying on the social network for information in real time. Users sometimes make decisions based on information posted by unfamiliar individuals on the social network [5] increasing the degree of reliance on the credibility of these sites. The Social network has succeeded in transforming the way different entities source and retrieve valuable information irrespective of their location. The Social network has also given users the privilege to give opinions with very little or no restriction.

Social Network – Power to the Users

Social sites have undoubtedly bestowed inconceivable privilege for their users to access the readily available never-ending uncensored information. For example:/ Twitter, permits its users to post events in real time way ahead of the broadcast of such events on traditional news media. The Social network also allows the user to express their views, i.e., be it may be positive or negative[4]. Organizations are now mindful of the implication of consumers' opinions posted on social network sites to the credit of their products or services and the overall success of their organizations. These entities follow the activities on the social network to keep side by side with how their audience reacts to issues that concern about them [4]. Considering the enormous volume of data being generated on the social network, it is important to find a computational means to filter, categorze, classify and analyze the social network contents.

Research Issues on Social Network Analysis

Some research issues that social networks face are as follows:

Linkage-Based and Structural Analysis

Link analysis is a data-analysis technique used to evaluate relationships between the nodes. Relationships was identified by among various type of nodes or objects, including organizations, people, and transactions. Link analysis has used by for investigation of criminal activity.

Dynamic Analysis and Static Analysis

Static analysis is assumed to be easier to carry out than streaming networks. Here the social network changes gradually over time and analysis on the entire network will be done in batch mode. Conversely, the dynamic analysis of streaming networks like Facebook and YouTube were very difficult to carry out than the former. Data generation using these networks are at high speed and capacity. The following sections present various data mining approaches used in analyzing social network data.

Graph-Based K-Means Clustering

Graph theory method is probably in the social network analysis in the early history of the social network concept. K-Means algorithm is the simplest and most commonly used vector quantization method. K-means clustering partitions data into clusters and minimizes the distance between cluster centers and data related to clusters. The approach was applied by the social network analysis to determine important features of the network such as the nodes and links.

Community Detection Using Hierarchical Clustering

The community is a smaller compressed group [7] within a larger network. Community formation is said to be one of the important characteristics of social networking websites. Community detection methods can assign a node not only to one community but also to many communities. Weights of all edges in complex networks are assumed to be the same in community detection methods. Communities on social networks, like any other communities in the real world, are very complex in nature and difficult to detect. Different authors have applied many clustering techniques to detect communities on the social network with hierarchical clustering is mostly used by authors.. Most hierarchical clustering methods require advance input. This technique is a combination of many techniques used to group nodes in the network to reveal the strength of individual groups which is then used to distribute the network into communities. Hierarchical clustering includes Vertex clustering method, where graph vertices was resolved by adding it in a vector space so that pair-wise length between vertices can be measured. Two peoples on the social network having several mutual friends are more likely to be closer than two people with fewer mutual friends in the network.

Semantic Web of Social Network

The Semantic Web environment makes knowledge sharing and reusability possible over different applications and community edges. Discovering the evolvement of Semantic Web (SW) enhances the knowledge of the importance of Semantic Web Community and emphasizes the synthesis of the Semantic Web. The work employs the concept of Friend of a Friend (FOAF) to explore how local and global community level groups expand and change in large-scale social networks on the *Semantic Web.* The study revealed the evolution outlines of social structures and forecasted future lift. Likewise application model of Semantic Web-based Social Network Analysis Model creates the ontological field library of social network analysis combined with the conventional outline of the semantic web to attain intelligent retrieval of the Web services.

Aspect-Based/Feature-Based Opinion Mining

Aspect-based also known as feature-based analysis is the process of mining the area of entity customers has reviewed. This is because not all aspects/features of an entity are often reviewed by customers. It is then necessary to summarise the aspects reviewed to determine the split of the overall review whether they are positive or negative. The sentiments expressed by some entities are easier to analyze than others, one of the reasons being that some reviews are ambiguous. The aspect-based opinion problem lies more in blogs and forum discussions than in product or service reviews. The aspect/entity (which may be a computer device) review is either 'thumb up', or 'thumb down', thumb up life form positive review while thumb down means review negative. Conversely, in blogs and forum discussions both aspects and entity are not familiar, and there are high levels of insignificant data which constitute noise. It is, therefore, necessary to identify opinion sentences in each review to

determine if indeed each opinion sentence is positive or negative. Opinion sentences can be used to summarize aspect-based opinion which enhances the overall mining of product or service review.

Sentiment Analysis of Social Network

Sentiment analysis also called as opinion mining. The main aim of it is to define the automatic tools able to extract one-sided information from texts in natural languages, such as opinions and sentiments, to create structured and actionable knowledge to be used by either a decision support system or a decision maker. Sentiment analysis was referred by discovery and recognition of positive or negative expression of opinion by people on diverse subject matters of interest. Depending on the field of application, several names is used sentiment analysis (e.g. opinion mining, opinion extraction, sentiment mining, subjectivity analysis, and review mining). It is commendable of note that the enormous opinions of several billions of the social network users are devastating; ranging from very important ones to mere assertions consequentially it has become necessary to examine sentiment expressed on social network with data mining techniques to generate meaningful frameworks was used by decision support tools. Diverse algorithms are in use to ascertain sentiment that matters to a topic, text, document or personality under review. The purpose of sentiment analysis on the social network is to recognize potential glide in the society as it concerns the attitudes, observations, and the expectations of stakeholder or the populace. This recognition enables the entities concern to take prompt actions by making necessary decisions. It is important to decode sentiment expressed to useful knowledge by way of mining and analysis.

Classification of Social Network Data

The Artificial neural network is the mathematical model based on the biological neural network. It consists of a set of processing units which communicate together by sending signals to each other over a large number of weighted connections. Such, processing units was called by neurons, and they are responsible for receiving input from neighbors or cells or variables or external sources and using this input to compute an output signal was propagated by other units. However, each unit also adjusts the weight of connections. It is very necessary to evolve neural network by modifying the weights of connections so that they become more accurate. The neural network should be trained by feeding it teaching patterns and letting it change its weights. This is learning process. There are three types of learning methods:

- Supervised learning where the network is trained by providing it with input and matching output patterns.
- Unsupervised learning where the output is trained to react to clusters of pattern within the input. There is no a priori set of categories into which the patterns are to be classified.
- Semi supervised learning where the test nodes need was predicted is known. ---- Reinforcement learning is the intermediate form between supervised and unsupervised learning.

Semi-Supervised Classification

Since the social network data usually come in huge sizes, in addition, also there are usually, a huge number of unlabelled instances. In such cases, it could be possible to use the information other than labels that exist in the unlabelled data, which leads to the use of semi-supervised learning algorithms. When the test nodes whose class will need to be predicted are known. In semi- supervised learning, the

unbalanced instances can be used to monitor the variance of the produces classifiers, to maximize the margin and hence to minimize the complexity.

Supervised Classification

While clustering techniques are used where the basis of data [8] was established but data pattern is unknown, classification techniques are supervised learning techniques used where the data organisation is already identified. It is worth of mentioning that understanding the problem to be solved and opting for the right data mining tool is very essential when using data mining techniques to solve social network issues. Pre-processing and considering privacy rights of individual should also be taken into account. Nonetheless, since social media is a dynamic platform, the collision of time can only be rational in the subject of topic recognition, but not substantial in the case of network enlargement, group behavior/ influence or marketing. This is because these attributes are bound to change from time to time.

Unsupervised Classification

In Data mining, unsupervised learning tries to the find hidden structure in unlabeled data. Since the examples given to the learner are unlabeled, then there is no error or reward signal to evaluate a potential solution.

Topic Detection and Tracking On Social Network

Topic Detection and Tracking *(TDT)* on social network employ different techniques for discovering the emergent of new topics [9] (or events) and for tracking their subsequent evolvements over a period. *TDT* is the delivery high level of attention recently. Many researchers and authors conduct research on *TDT* on social network sites, especially on Twitter [5]; [6]. The main aim of TDT was to develop core technologies for news understanding systems. More specifically its tasks focused on discovering and keeping track of real-world events in multi-lingual news streams from various sources. Various methods have been residential for this task, including machine learning and query expansion based methods.

Conclusion and Future Work

Different data mining techniques are used to the social network analysis as covered in this work. The techniques are ranging from unsupervised to semi-supervised and supervised learning methods. As far, different levels of improvements have been achieved either with combined or solitary techniques. The result of the experiments conducted on social network analysis is supposed to have shed more light on the structure and activities of social networks. The varied experimental results have also established the bearing of data mining techniques in retrieve valuable information and contents from huge data generated on social network. The future survey will tend to investigate novel up to date data mining techniques for social network analysis. The survey will compare parallel data mining tools and suggest the most fitting tool(s) for the dataset to be analyzed. The table also contains the approaches engaged, the experimental results and the dates and authors of the approaches.

References

1. Borgatti, S P.: "2-Mode concepts in social network analysis." Encyclopedia of Complexity and System Science, 8279-8291, 2009.

2. Asur, S., and Huberman, B.: "Predicting the future with the social network." Web Intelligence Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on. Vol. 1. IEEE, 2010.
3. Aggarwal, C.: An introduction to social network data analytics. Springer US, 2011.
4. Castellanos, M., Dayal, M., Hsu, M., Ghosh, R., Dekhil, M.: U LCI: A Social Channel Analysis Platform for Live Customer Intelligence. In: Proceedings of the 2011 International Conference on Management of Data. 2011.
5. Pang, B. and Lee, L.: Opinion mining and sentiment analysis; Foundations and trends in information Retrieval; Vol. 2, Nos. 1–2, 1–135, 2008.
6. Adedoyin-Olowe, M., Gaber, M., Stahl, F.: A Methodology for Temporal Analysis of Evolving Concepts in Twitter. Proceedings of the 2013 ICAISC, International Conference on Artificial Intelligence and soft computing. 2013.
7. Becker, H., Naaman, M., Gravano, L.: Beyond Trending Topics: Real-World Event Identification on Twitter. ICWSM, 11, 438-441, 2011.
8. Adomavicius, G., Tuzhilin, A.: Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. Knowledge and Data Engineering, IEEE Transactions on, 17(6), 734-749, 2005.
9. Https://www.researchgate.net/publication/259335258_A_Survey_of_Data_Mining_Techniques_for_Social_Media_Analysis [accessed Jan 28 2018].

A REVIEW ON BIG DATA RECOMMENDATION SYSTEM

R.Ilakkiya
M.Phil. Scholar, Department of Computer Science
Raja Doraisingam Government Arts College, Sivaganga, TamilNadu, India

A.Prema
Assistant Professor, Department of Computer Science
Raja Doraisingam Government Arts College, Sivaganga, TamilNadu, India

N.Sujatha
Assistant Professor, Department of Computer Science
Sri Meenakshi Government Arts College for Women, Madurai, Tamilnadu, India

Abstract
Now a day's online searching process increases and people searches new information in the search process. Recommender system involves in the process and implements as services. Recommendation Engines are one of the easiest areas of Big data. This operation is a very high CPU consuming task and it runs for hours to make it parallel and fast, for this solutions like Hadoop framework can be used .The proposed work analyze occurring issues collaborative filtering Algorithm generates keyword recommenders from the previous user preferences.
Keywords*: Big data, Recommendation system, Collaborative filtering, Hadoop tool*

Introduction

Big data is a blanket term for the non-traditional strategies and technologies needed to gather, organize, process, and gather insights from large datasets. While the problem of working with data that exceeds the computing power or storage of a single computer is not new, the pervasiveness, scale, and value of this type of computing has greatly expanded in recent years. Big data is a "large dataset" that means a dataset too large to reasonably process or store with traditional tooling or on a single computer. This means that the common scale of big datasets [4] is constantly shifting and may vary significantly from organization to organization. These datasets can be orders of magnitude larger than traditional datasets, which demands more thought at each stage of the processing and storage life cycle. Another way in which big data differs significantly from other data systems is the speed that information moves through the system. Data is frequently flowing into the system from multiple sources and is often expected to be processed in real time to gain insights and update the current understanding of the system.

Big data problems are often unique because of the wide range of both the sources being processed and their relative quality. Data can be ingested from internal systems like application and server logs, from social media feeds and other external APIs [2, 6], from physical device sensors, and from other providers. Big data seeks to handle potentially useful data regardless of where it's coming from by consolidating all information into a single system.

Big data also brings new opportunities and critical challenges to industry and academia. Similar to most big data applications, the big data development also positions heavy impacts on service recommender systems. With the number of alternative services, effectively recommending services [1, 3, and 7] that users preferred have become an important research issue. Service recommender systems

provided valuable tools to help users deal with services overload and appropriate recommendations. Recommender systems apply techniques and methodologies from another neighboring areas -such as Human computer Interaction (HCI)or Information Retrieval(IR) [8,10,11]. However, most of these systems bear in their core an algorithm that can be understood as a particular instance of a data mining (DM) technique. Recommender System makes use of different sources of information for providing users with predictions and recommendations of items. They try to balance factors like accuracy, novelty, disparity and stability in the recommendations. Collaborative Filtering (CF) methods play an important role in the recommendation, although they are often used along with other filterning techniques like content-based, knowledge-based or social ones. CF is based on the way in which humans have made decisions throughout history besides on our own experiences; we also base our decisions on the experiences and knowledge that reach each of us from a relatively large group of acquaintances.

Methodologies in Recommender System

The process for generating RS recommendation is based on a combination of the following considerations:

- The type of data available in its database (e.g., ratings, user registration information, features and content for items that can be ranked, social relationships among users and location-aware information).
- The filtering algorithm used (e.g., demographic, content-based, collaborative, social-based, context-aware and hybrid). The model chosen (e.g., based on direct use of data: ‘‘memory based,’’ or a model generated using such data: ‘‘model-based’’).
- The employed techniques are also considered probabilistic approaches, Bayesian networks, nearest neighbor’s algorithm bio-inspired algorithms such as neural networks and genetic algorithms fuzzy models, singular value decomposition techniques to reduce sparsity levels, etc.
- Sparsity level of the database and the desired scalability.
- Performance of the system (time and memory consuming).
- The objective sought is considered (e.g., predictions and top N recommendations) as well as the desired quality of the results (e.g., novelty, coverage and precision). Resear

Content-Based Filtering

Content-based filtering makes recommendations based on user choices made in the past (e.g. in a web-based e-commerce RS, if the user purchased some fiction films in the past, the RS will probably recommend a recent fiction film that he has not yet purchased on this website). Content-based filtering [10, 13, and 15] also generates recommendations using the content from objects intended for recommendation therefore; certain content can be analyzed, like text, images and sound. From this analysis, a similarity can be established between objects as the basis for recommending items similar to items that a user has bought, visited, heard, viewed and ranked positively.

Demographic filtering

Demographic filtering is justified on the principle that individuals with certain common personal attributes [9, 16] (sex, age, country, etc.) will also have common preferences.

Collaborative Filtering

Collaborative Filtering allows users to give ratings about a set of elements (e.g. videos, songs, films, etc. in a CF based website) in such a way that when enough information is stored on the system, we can make recommendations[5,7] to each user based on information provided by those users we consider to

have the most in common with them. CF is an interesting open research field. As noted earlier, user ratings can also be implicitly acquired (e.g., number of times a song is heard, information consulted and access to a resource).

Hybrid Filtering

Hybrid filtering commonly uses a combination of CF with demographic filtering or CF with content-based filtering to exploit merits of each one of these techniques. Hybrid filtering is usually based on bioinspired or probabilistic methods such as genetic algorithms, fuzzy genetic, neural networks, Bayesian networks, clustering and latent features.

Memory based Methods

Memory-based methods can be defined as methods that (a) act only on the matrix of user ratings for items and (b) use any rating generated before the referral process (i.e., its results are always updated). Memory-based methods usually use similarity metrics to obtain the distance between two users, or two items, based on each of their ratios.

Model based methods

Model-based methods use RS information to create a model that generates the recommendations. Herein, we consider a method model-based if new information from any user outdates the model. Among the most widely used models we have Bayesian classifiers, neural networks, fuzzy systems, genetic algorithms, latent features and matrix factorization, among others.

Metrics in Recommender Systems

Research in the RS field requires quality measures and evaluation metrics to know the quality of the techniques, methods, and algorithms for predictions and recommendations. Evaluation metrics and evaluation frameworks facilitate comparisons of several solutions for the same problem and selection from different promising lines of research that generate better results.

Evaluation metrics can be classified as (a) prediction metrics such as the accuracy ones Mean Absolute Error (MAE), Root of Mean Square Error (RMSE), Normalized Mean Average Error (NMAE); and the coverage (b) set recommendation metrics such as Precision, Recall and Receiver Operating Characteristic (ROC)] (c) rank recommendation metrics such as the half-life and the discounted cumulative gain] and (d) diversity metrics: such as the diversity and the novelty of the recommended items. The validation process is performed by employing the most common cross validation techniques (random sub-sampling and k-fold cross validation) for cold-start situations, due to the limited number of users (or items) votes involved, the usual method chosen to carry out the experiments is leave-one out cross validation.

In order to measure the accuracy of the results of an RS, it is usual to use the calculation of some of the most common prediction error metrics, amongst which the Mean Absolute Error (MAE) and its related metrics: mean squared error, root mean squared error, and normalized mean absolute error stand out.

The novelty evaluation measure indicates the degree of difference between the items recommended to and known by the user. The diversity quality measure indicates the degree of differentiation among recommended items.

The stability in the predictions and recommendations influences on the users trust towards the RS. A RS is stable if the predictions provide do not change strongly over a short period of time.

Comparative Analysis

Metrics	CB Filtering	Demographic filtering	Collaborative Filtering	Hybrid Filtering
Methodology	Content based RS like text, images, video	Filter based on user's personal details	Filter data based on user's ratings	Combine CB and Demographic filtering
Accuracy	More accurate	Normal	Accurate results	Predict accurate results
Speed	Low	High	High	Normal
Scalability	Increases when content increases	Increases when combine more factors	Does not affect when content increases	Affects when combines filtering methods
Memory usage	High	Medium	High	Medium
Prediction algorithms	More algorithms	Limited	More algorithms	Normal

Conclusion

Recommender system involves in this process and implements as service. Service recommender system gives additional information to the user but if information grows then these process become a critical one. The proposed work analyses issues occurring when service recommender system implements in large data sets. This work proposes a keyword-Aware services Recommender method, to split the services to the users and mainly focused keywords from the user preferences. This paper analyses recommender system methodologies and evaluates its metrics. Collaborative recommender system predicts relevant results from user's data and reduces search time. This method has greater flexibility compared to other methods and achieves optimal results.

References

1. Michael J. Pazzani and Daniel Billsus, "Content-based recommendation systems," Springer Berlin Heidelberg. pp. 325-341,2007
2. Badrul Sarwar, George Karypis, Joseph Konstan, and John Ried, "Item-based Collaborative Filtering Recommendation Algorithms," May 1-5, 2001, Hong Kong.
3. R. Burke, "Hybrid Recommender Systems: Survey and Experiments," User Modeling and User-Adapted Interaction, vol. 12, no. 4, pp. 331-370, 2002.
4. L. Sharma and A. Gera, "A Survey of Recommendation System: Research Challenges," International Journal of Engineering Trends and Technology (IJETT), vol. 4, May 2013.
5. Shunmei Meng, Wanchun Dou, Xuyun Zhang,"KASR: A Keyword-Aware Service Recommendation Method on Mapreduce for Big Data Applications," IEEE Trans. on Parallel and Distributing systems, vol.25, no.12, December 2014.
6. Y. Chen, A. Cheng, and W. Hsu, "Travel Recommendation by Mining People Attributes and Travel Group Types from Community-Contributed Photos," IEEE Trans. Multimedia, vol. 25, no. 6, pp. 1283-1295, Oct. 2013
7. M. Alduan, F. Alvarez, J. Menendez, and O. Baez, "Recommender System for Sport Videos Based on User Audiovisual Consumption," IEEE Trans. Multimedia, vol. 14, no. 6, pp. 1546-1557,Dec. 2012.

8. S. Alonso, F.J. Cabrerizo, F. Chiclana, F. Herrera, E. Herrera-Viedma, Group decision making with incomplete fuzzy linguistic preference relations, International Journal of Intelligent Systems 24 (2009) 201–222.
9. Ansari, S. Essegaier, R. Kohli, Internet recommendation systems, Journal of Marketing Research 37 (3) (2000) 363–375.
10. N. Antonopoulus, J. Salter, Cinema screen recommender agent: combining collaborative and content-based filtering, IEEE Intelligent Systems (2006) 35– 41.
11. P. Antunes, V. Herskovic, S.F. Ochoa, J.A. Pino, Structuring dimensions for collaborative systems evaluation, ACM Computing Surveys 44 (2) (2012). Article 8.
12. O. Arazy, N. Kumar, B. Shapira, Improving Social Recommender Systems, Journal IT Professional 11 (4) (2009) 31–37.
13. L. Ardissono, A. Goy, G. Petrone, M. Segnan, P. Torasso, INTRIGUE: Personalized recommendation of tourist attractions for desktop and handset devices, Applied Artificial Intelligence 17 (8-9) (2003) 687–714.
14. R. Baeza-Yates, B. Ribeiro-Neto, Modern Information Retrieval, AddisonWesley, 1999.
15. M. Balabanovic, Y. Shoham, Content-based, collaborative recommendation, Communications of the ACM 40 (3) (1997) 66–72.
16. L. Baltrunas, T. Makcinskas, F. Ricci, Group recommendation with rank aggregation and collaborative filtering, in: Proceedings of the 2010 ACM Conference on Recommender Systems, 2010, pp. 119–126

SCRUTINIZE OF MASSIVE DATA AND IOT

P.Rohini
Student, Department of Computer Science
Raja Doraisingam Government Arts College Sivaganga, Tamilnadu, India

A.Prema
Assistant Professor, Department of Computer Science
Raja Doraisingam Government Arts College, Sivaganga, Tamilnadu, India

K.Chelladurai
Assistant Professor, Department of Computer Science
Raja Doraisingam Government Arts College, Sivaganga, Tamilnadu, India

Abstract
Big data refers to data sets that are not only big but also high in range and rate which makes them complex to handle using common toolsand technique. Big data demonstrates the $10v_s$ volume, variety, velocity, veracity, values, validity, variability, vocabulary and vagueness. The internet of things(IOT) can be described as connecting everyday objects like smart-phones, internet TVs, sensorsand actuators to the internet where the device are intelligently linkedtogether enabling new forms of communications between things and people and between things themselves. IOT is aggregating and compressing massive amounts of low latency machine –generated data. IOT aims to takes a wide range of "things" and turn them into smart objects. Hadoop is on open first place software project that enables the freckled processing of huge data sets across group of expose of trade service. Hadoop is very much suitable for high volume of data and it also provide the high speed access to the data of the application which want to use. In this paper we will briefly introduce the massive data and IOT.
Keywords: *Big data, IOT (Internet of Things), Hadoop.*

Introduction

Big data means actually a big data; it is a collection of huge datasets that cannot be handled using old computing techniques. Big data is not only containing data, it also contains various tools, techniques and frameworks. Data that has extra-large Volume, comes from Variety of sources, Variety of formats and comes at us with a great Velocity is normally referred to as Big Data.The data in it will be of three types.

Structured data: Relational data. Semi Structured data: XML data.Formless data: Word, PDF, Text, Media Logs.

The Internet of Things (IOT) is a work of fictionmodel that is quickly gaining ground in the scenario of modern wireless telecommunications. The fundamental idea of this concept is the invasiveincidence around us of a range of things or objects – such as Radio-Frequency IDentification (RFID) tags, sensors, actuators, mobile phones, etc.

Big Data Parameters

As the data is bigger from different sources in different form, it is represented by the 10 Vs. [1]

Volume

Volume means ratio of data or huge amount of data develop in every second. Machine develop data are examples for these components. Nowadays data volume is increasing from gigabytes to peta bytes. 40 Zetta bytes of data will be created by 2020 which is 300 times from2005.

Velocity

Velocity is the speed at which data is developing and processed. For example social media posts.

Variety

Variety is one more important characteristic of big data. It refers to the type of data. Data may be in different styles such as Text, numerical, images, audio, video data. On twitter 400 million tweets are sent per day and there are 200 million active users on it.

Variably

Variability in bi data's context refers to a few different things. Big data is also variable because of the multitude of data dimensions resulting from multiple disparate data types and sources.

Veracity

Veracity means anxiety or accuracy of data. Data is uncertain due to the inconsistency and in completeness.

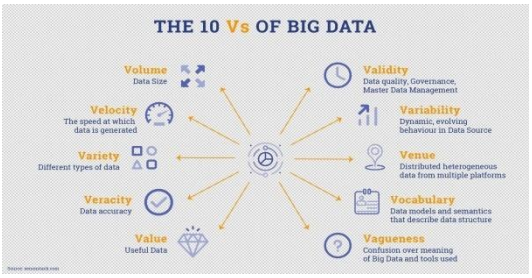


Figure 1:10 v's in big data [8]

Value

The all-important v, characterizing the business value, and potential of big data to transform our organization from top to bottom.

Validity

Similar to reliability, validity refers to how correct and spot-on the data is for its intended use. The benefit from hugedata analytics is only as good as its underlying data, common definitions, and metadata.

Venue

Circulated heterogeneous data from many platforms, with different access and formatting requirements.

Vocabulary

Scheme, data models, semantics, and other content-and context-based metadata that describe the data's structure.

Vagueness

Confusion over the meaning of big data.

Technologies and Methods [1]

All paragraphs must be indented. Big data is a new concept for handling massive data therefore the architectural description of this technology is very new. There are different technologies which use almost same approach i.e. to distribute the data among various local agents and reduce the load of the main server, so that traffic can be avoided. There are endless articles, books and periodicals that describe Big Data from a technology perspective so we will instead focus our efforts here on setting out some basic principles and the minimum technology foundation to help relate Big Data to the broader IM domain.

A. Hadoop

Hadoop is a framework that can run applications on systems with thousands of nodes and terabytes. It distributes the file among the nodes and allows system to continue work in case of a node failure. This approach reduces the risk of catastrophic system failure.

The application is broken into smaller parts (fragments or blocks).Apache Hadoop consists of the Hadoop kernel, Hadoop distributed file system (HDFS), map reduce and related projects are zookeeper, Hbase, Apache Hive. Hadoop Distributed File System consists of three Components: the Name Node, Secondary Name Node and Data Node. The multilevel secure (MLS) environmental problems of Hadoop by using security enhanced Linux (SE Linux) protocol. In which multiple sources of Hadoop applications run at different levels.

This protocol is an extension of Hadoop distributed file system. Hadoop is commonly used for distributed batch index building. Iit is desirable to optimize the index capability in near real time. Hadoop provides components for storage and analysis for large scale processing.

The advantage of Hadoop is Distributed storage & Computational capabilities, extremely scalable, ptimized for high throughput, large block sizes, tolerant of software and hardware failure.

B. Map Reduce Components

1. **Name Node**: manages HDFS metadata, doesn'tdeal with files directly.
2. **Data Node**: stores blocks of HDFS—defaults replication level for each block: 3.
3. **Job Tracker**: schedules, allocates and monitorsjob execution on slaves—Task Trackers.
4. **Task Tracker**: runs Map Reduce operations.[2]

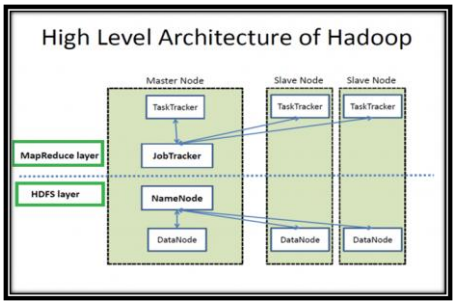


Figure 2: Hadoop architecture [11]

C.HDFS

The Hadoop Distributed File System (HDFS) is the file system component of the Hadoop framework. HDFS is designed and optimized to store data over a large amount of low-cost hardware in a distributed fashion.

- **NameNode:** The NameNode is the central location for information about the file system deployed in a Hadoop environment. An environment can have one or two NameNodes, configured to provide minimal redundancy between the NameNodes. The NameNode is contacted by clients of the Hadoop Distributed File System (HDFS) to locate information within the file system and provide updates for data they have added, moved, manipulated, or deleted.
- **DataNode:**DataNodes make up the majority of the servers contained in a Hadoop environment. The DataNode serves two functions It contain a portion of the data in the HDFS and it acts as a compute stage for running jobs, some of which will develop the local data within the HDFS.

The Important Basis Behind Why To Implement Iot And Big Data Are: [7]

1. Analytical monitoring
2. More Uptime
3. Lower reject rates
4. Higher throughput
5. Enhanced safety
6. Efficient use of labor
7. Enable mass customization
8. Analyze the activities for real-time marketing
9. Improved situational alertness

10. Improved quality
11. Sensor-driven decision analytics
12. Process optimization
13. Optimized resource utilization
14. Instant control and response in complex independent systems.

The above mentioned are some possible reasons to implement IOT and Big data. As the requirements of both the technologies go hand in hand, a proper improved system is needed to overcome the challenges they pose. Many companies strive to meet the challenges and take possible steps to overcome them.

Internet of Things Generates Big Data

The 'Internet of Things' can generate 'Big Data' for a number of reasons [9]. The volume of data attributable to the 'Internet of Things' is substantial. As sensors interact with the world, 'Things'such as RFID tags generates volumes and volumes of data. As a result, digital processing becomes a requirement of feasibility. The velocity of data associated with the 'Internet of Things', compared with traditional transaction processing, explodes as sensors can continuously capture data. The variety of data associated with the 'Internet of Things' also is expansive as the types of sensorsand the different sources of data expand. The veracity of data in the 'Internet of Things' may also beimproving as the quality of sensor and other data improves over time. For example, use of RFID tagsgenerates much more reliable information than a decade ago. Such high volumes of data, coupled withan increasing velocity of data, along with an increased variety of data, illustrate the push by the 'Internetof Things' to generate 'Big Data'.

Application of Iot Based Big Data

IOT (Internet of Things) is not only an important source of big data, but also one of the main markets of big data applications.[10] Because of the high variety of objects, the applications of IOT also evolve endlessly. Logistic enterprises may have profoundly experienced with the application of IOT big data. The application of smart city brings aboutBenefits in many aspects for Dade County. For instance, theDepartment of Park Management of Dade County saved onemillion USD in water bills due to timely identifying andfixing water pipes that were running and leaking this year.

Challenges of Iot Big Data

Major challenges that can fetch momentous rewards when they are solved [7].

1. Huge data volumes
2. Difficulty in data collection
3. Incompatible standards
4. New security threats
5. No reliability in the data
6. Fundamental shifts in business models
7. Huge amount of data to analyze
8. A rapidly evolving privacy landscape

The above points are some of the challenges that IOT big data faces. The rate in data growth is expanding every second, storage in a big challenge, processing and maintaining is even more tedious. The tools that are developed to manage the both technologies are day by day changing as per the requirements. No doubt, both technologies are going to play a major role in the information technology field.

Conclusion

This paper is a review about the IOT and Big data, the impacts of IOT on big data, the big data technologies and the challenges. Since there is a major impact of IOT on big data we need to manage quicklythe complete structure to handle the daily changing situationthere are a few areas of concern and security and privacy and data collection efficiency are probably the most difficult problems we are facing. Security compromise and inefficiencies in data collection mechanisms result in a loss of status, money, time and effort. But there is hope because both the IOT and the big data are at apromising stage and there will be improve.

References

1. Review Paper On Big Data Using Hadoop Ms. GurpreetKaur Student, PG Department of Computer Science, Mata GujriCollege Fatehgarh Sahib, Punjab, Ms. ManpreetKaur Assistant Professor, PG Department of Computer Science, Mata Gujri College, Fatehgarh Sahib, Punjab
2. Survey Paper on Big Data C. Lakshmi*, V. V. Nagendra Kumar MCA Department, RGMCET, Nandyal, Andhra Pradesh, India, Volume 6, Issue 8, August 2016 ISSN: 2277 128X
3. IoT and Big Data- The Current and Future Technologies: A Review, K.R.Kundhavai1, S.Sridevi2 1Assistant Professor, 2Assistant Professor 1New Horizon College of Engineering, 2New Horizon College of Engineering Bangalore, India
4. The Internet of Things: A survey, Luigi Atzori a, Antonio Iera b, GiacomoMorabitoc,*a DIEE, University of Cagliari, Italy b University ‘‘Mediterranea” of Reggio Calabria, Italy cUniversity of Catania, Italy
5. Considerations for Web of Object Service Architecture on IoT Environment,NamKyung Lee*, HyunWoo Lee* and Won Ryu* * Electronics and Telecommunications Research Institute nklee@etri.re.kr
6. Research Issues in Big Data Analytics Manish Kumar Kakhani1, Sweeti Kakhani2 and S.R. Biradar3 1Assistant Professor, Faculty of Engineering and Technology, MITS University, Lakshmangarh, Rajasthan 2 Lecturer, Faculty of Arts, Science and Commerce, MITS University, Lakshmangarh, Rajasthan 3 Professor, Faculty of Engineering and Technology, MITS University, Lakshmangarh, Rajasthan
7. K.R.Kundhavai*et al*, International Journal of Computer Science and Mobile Computing, Vol.5 Issue.1, January- 2016, pg. 10-14
8. http://goo.gl/iagges/jR443f
9. ‘Big Data’, The‘Internet Of Things’ And The ‘InternetOf Signs’, Daniel E. O’leary*, University of Southern California, Los Angeles, CA, USA
10. Big Data: A Survey, Min Chen · Shiwen Mao · Yunhao Liu, © Springer Science+Business Media New York 2014
11. Magnai17, hadoop-HighLevel_hadoop_architecture-640*460.png, 21 march 2016.
12. Please cite this article as: E. Ahmed et al., The role of big data analytics in Internet of Things, Computer Networks (2017), http://dx.doi.org/10.1016/j.comnet.2017.06.013
13. International Journal of Computer Engineering & Technology (IJCET)
14. Volume 6, Issue 12, Dec 2015, pp. 65-71, Article ID: IJCET_06_12_008
15. Available online at
16. http://www.iaeme.com/IJCET/issues.asp?JType=IJCET&VType=6&IType=12

ANDROID APP FOR WEBSITE

A.Prema
Assistant Professor, Department of Computer Science
Raja Dorai Singam Government Arts College, Sivagangai, TamilNadu, India

N.Sujatha
Assistant Professor, Department of Computer Science
Sri Meenakshi Govt Arts College for Women, Madurai Tamilnadu, India

M.Sujithra
PG Student, Department of Computer Science
Raja Dorai Singam Government Arts College, Sivagangai, TamilNadu, India

Abstract
In topical years, the emergence of smart phones has changed the description of mobile Phones . Phone is no longer just a contact tool, but also an essential part of the people's communication and daily life. Various applications added boundless fun for people's lives. It is firm that the future of the network will be the mobile life-threatening. Now the Android system in the electronics market is becoming more and more fashionable, especially in the smart phone market. Because of the open source, some of the expansion tools are free, so there are plenty of applications generated. This drastically stirred the people to use the Android system. In addition, it provides a very handy hardware platform for developers so that they can spend less effort to grasp their ideas. This makes Android can get auxiliary growth. We propose the Android App for the website with responsive layout. In short we will be using them to complete our daily tasks. One application that cascade into this category is the Android App for Website developed for the android mobiles.
Keywords: *Android, Mobile, Website, App.*

Introduction

"Android app for Website" is the one app you just sort of always expect to be there, which is often a good thing and a bad thing. Everyone knows accessing the web is a critical part of almost every device today, but a surprising number of users rarely use something other than the browser that was included with their phone.

As an Android user, there's a good chance you already know about some website. It probably came pre-loaded on your phone, unless you've had your Android phone for a little while or you just didn't notice it. website is a powerful extension of website for Windows and OSX, and one of its best features is the ability to sync your browser tabs from one device to another, so you can be looking something up on your laptop and pull it up on your phone when you're out and about. If you're an Android 5.0+ user, Chrome for Android can also split out its tabs as individual windows so they look like their own apps in the app switching interface.

It's also pretty great at navigating the web, offering a great mix of fast mobile-friendly experiences and full desktop browser capabilities depending on the needs at hand, and includes settings for data conservation and an incognito mode for browsing sites in which you need not to be tracked or logged through. As far as decent all-around browsers go, you're unlikely to need too much more than our website.

To create the Android App for our college website. Because when we need to see our college website we need some web browser and browse our web site it will take lots of time as well as waste of time. If somebody wants to know about.

Website its not necessary to have a internet connection. If we are having Bluetooth we can transfer the apps and we can share it. This apps will fit in all the device which contain the full responsive layout.

About how to build apps using Android's various APIs. Android provides a rich application framework that allows you to build innovative apps and games for mobile devices in a Java language environment. The documents listed in the left navigation provide details

Apps provide multiple entry points

Android apps are built as a combination of distinct components that can be invoked individually. For instance, an individual *activity* provides a single screen for a user interface, and a *service* independently performs work in the background.

From one component you can start another component using an *intent*. You can even start a component in a different app, such as an activity in a maps app to show an address. This model provides multiple entry points for a single app and allows any app to behave as a user's "default" for an action that other apps may invoke.

Responsive layout

Android provides an adaptive app framework that allows you to provide unique resources for different device configurations. For example, you can create different XML layout files for different screen sizes and the system determines which layout to apply based on the current device's screen size.

You can query the availability of device features at runtime if any app features require specific hardware such as a camera. If necessary, you can also declare features your app requires so app markets such as Google Play Store do not allow installation on devices that do not support that feature.

Resources Overview

Our should always externalize resources such as images and strings from your application code, so that you can maintain them independently. Externalizing your resources also allows you to provide alternative resources that support specific device configurations such as different languages or screen sizes, which becomes increasingly important as more Android-powered devices become available with different configurations. In order to provide compatibility with different configurations, you must organize resources in your project's res/ directory, using various sub-directories that group resources by type and configuration.

For any type of resource, you can specify *default* and multiple *alternative* resources for your application:

- Default resources are those that should be used regardless of the device configuration or when there are no alternative resources that match the current configuration.
- Alternative resources are those that you've designed for use with a specific configuration. To specify that a group of resources are for a specific configuration, append an appropriate configuration qualifier to the directory name.
- For example, while your default UI layout is saved in the res/layout/ directory, you might specify a different layout to be used when the screen is in landscape orientation, by saving it in the res/layout-land/ directory. Android automatically applies the appropriate resources by matching the device's current configuration to your resource directory names.

He following documents provide a complete guide to how you can organize your application resources, specify alternative resources, access them in your application, and more:

Providing Resources

What kinds of resources you can provide in your app, where to save them, and how to create alternative resources for specific device configurations.

Accessing Resources

How to use the resources you've provided, either by referencing them from your application code or from other XML resources.

Handling Runtime Changes

How to manage configuration changes that occur while your Activity is running.

A bottom-up guide to localizing your application using alternative resources. While this is just one specific use of alternative resources, it is very important in order to reach more users.

Resource Types

A reference of various resource types you can provide, describing their XML elements, attributes, and syntax. For example, this reference shows you how to create a resource for application menus, drawables, animations, and more.

Working with XML on Android

Android is a modern, open source operating system and SDK for mobile devices. With ityou can create powerful mobile applications. This becomes even more attractive when yourapplications can access Web services, which means you need to speak the language of theWeb: XML. In this article, you will see different options for working with XML on Android andhow to use them to build your own Android applications.

In this article, you learn to build Android applications that can work with XML from the Internet.Android applications are written in the Java™ programming language, so experience with Javatechnology is a must-have. To develop for Android, you will need the Android SDK. All of the codeshown in this article will work with any version of the Android SDK, but SDK 1.5_pre was usedto develop the code. You can develop Android applications with just the SDK and a text editor,but it is much easier to use the Android Developer Tools (ADT), an Eclipse plugin. For this article,version 0.9 of ADT was used with Eclipse 3.4.2, Java edition. See Resources for links to all ofthese tools.

XML on Android

The Android platform is an open source mobile development platform. It gives you access to allaspects of the mobile device that it runs on, from low level graphics, to hardware like the cameraon a phone. With so many things possible using Android, you might wonder why you need tobother with XML. It is not that working with XML is so interesting; it is working with the things that it enables. XML is commonly used as a data format on the Internet. If you want to access datafrom the Internet, chances are that the data will be in the form of XML. If you want to send data toa Web service, you might also need to send XML. In short, if your Android application will leveragethe Internet, then you will probably need to work with XML. Luckily, you have a lot of optionsavailable for working with XML on Android.

XML parsers

Frequently used acronyms

API: Application programming interface
RSS: Really Simple Syndication
SDK: Software Developers Kit
UI: User interface
URL: Universal Resource Locator
XML: Extensible Markup Language

One of the greatest strengths of the Android platform is that it leverages the Java programminglanguage. The Android SDK does not quite offer everything available to your standard JavaRuntime Environment (JRE,) but it supports a very significant fraction of it. The Java platform hassupported many different ways to work with XML for quite some time, and most of Java's XML-related APIs are fully supported on Android. For example, Java's Simple API for XML (SAX) andthe Document Object Model (DOM) are both available on Android. Both of these APIs have beenpart of Java technology for many years. The newer Streaming API for XML (StAX) is not availablein Android. However, Android provides a functionally equivalent library. Finally, the Java XMLBinding API is also not available in Android. This API could surely be implemented in Android.However, it tends to be a heavyweight API, with many instances of many different classes oftenneeded to represent an XML document. Thus, it is less than ideal for a constrained environmentsuch as the handheld devices that Android is designed to run on. In the following sections, you willtake a simple source of XML available on the Internet, and see how to parse it within an Androidapplication using the various APIs mentioned above. First, look at the essential parts of the simpleapplication that will use XML from the Internet.

App Manifest

Every application must have an AndroidManifest.xml file (with precisely that name) in its root directory. The manifest file presents essential information about your app to the Android system, information the system must have before it can run any of the app's code. Among other things, the manifest does the following:

- It names the Java package for the application. The package name serves as a unique identifier for the application.
- It describes the components of the application — the activities, services, broadcast receivers, and content providers that the application is composed of. It names the classes that implement each of the components and publishes their capabilities (for example, which Intent messages they can handle). These declarations let the Android system know what the components are and under what conditions they can be launched.
- It determines which processes will host application components.
- It declares which permissions the application must have in order to access protected parts of the API and interact with other applications.
- It also declares the permissions that others are required to have in order to interact with the application's components.
- It lists the Instrumentation classes that provide profiling and other information as the application is running. These declarations are present in the manifest only while the application is being developed and tested; they're removed before the application is published.
- It declares the minimum level of the Android API that the application requires.
- It lists the libraries that the application must be linked against.

Proposed Work

Allows full use of the device and API. No programming limitations, anything and everything is possible (that the device allows) well documented for custom requirements e.g. custom maps. Runs faster and smoother than other methods.

Less likelihood for bugs. Less likelihood for yearly updates. Updates are quick if required. guaranteed to run on the device it has been built for and look as intended can be sold via iTunes, Android Market etc. Quicker to develop than native apps.

Designed to work with apps that use. camera, gps, text/image display, accelerometer, google/apple maps over internet connection. (most) code can be used across platforms Android/iOs etc. cheaper to develop.

Looks like a native app. can still be sold on iTunes, Android Market etc.data can be stored on a web server with a Content Management System allowing realtime updates to app content (web site is updated and so is app at the same time). allows for using google/apple maps on device (simple to plot routes, points of interest, familiar interface).

Cheaper in the long run if programmer doesn't have to do content updates for client.

You can make your web content available to users in two ways:

1. A traditional web browser.
2. Android application, by including a Web View in the layout.

There are essentially two ways to deliver an application on Android:

1. As a client-side application (developed using the Android SDK and installed on user devices in an APK).
2. Web application (developed using web standards and accessed through a web browser—there's nothing to install on user devices).

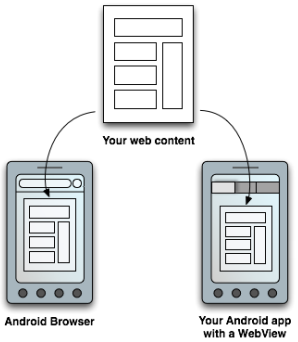


Figure 1 : Process of the proposed Android App

If you choose to provide a web-based app for Android-powered devices, you can rest assured that major web browsers for Android (and the Web View framework) which allow you to specify viewport and style properties that make your web pages appear at the proper size and scale on all screen configurations.

Illustrates how you can provide access to your web pages from either a web browser or your own Android app. However, you shouldn't develop an Android app simply as a means to view your web site. Rather, the web pages you embed in your Android app should be designed especially for that environment. You can even define an interface between your Android application and your web pages that allows JavaScript in the web pages to call upon APIs in your Android application—providing Android APIs to your web-based application.

Conclusion

The advantages of Responsive Web Design include a single code base that provides easy and low maintenance along with a single version of the website that improves SEO. Mobile Web App provides a better user experience, lower risk of performance issues and faster implementation time for your mobile service if you know your exact requirements.

Therefore, if you want a solution that's easy to maintain, makes use of existing skills and that you can control, then Responsive Web Design is the approach for you. If you want a high quality user

experience, better performance, faster implementation and seamless transactions, then you should select Mobile Web App for our Organization.

References

1. Android Programming: The Big Nerd Ranch Guide (Big Nerd Ranch Guides) Kindle Edition by Brian Hardy
2. Very Useful Android Java Code For Snippets for Beginners By Andrei Dan

Web References

1. http://www.Androidcentral.com
2. http://www.w3schools.com/php/default.asp
3. http://www.joomlatutorials.com/
4. http://www.tutorialspoint.com/mysql/
5. http://www.developer.android.com/
6. http://www.rapidvaluesolution.com/
7. http://www.Androidtutorial.com/
8. http://www.Androidschemas.com/
9. http://www.luna.eclipse.com/
10. https://www.quora.com/What-is-the-best-mobile-browser-for-Android
11. http://www.android.eclipse.org/releases/luna